GENERATING MULTIVARIATE NORMAL DATA IN SAS

Daniel M. Chilko and E. James Harner
West Virginia University
Morgantown, West Virginia

ABSTRACT


Computer random number generators offer a solution to the problem of finding suitable examples of data for classroom use. A SAS procedure for generating multivariate normal data has been developed. This provides the instructor a single system to generate data, analyze them, and display results. It is used in teaching multivariate statistics.

INTRODUCTION

The use of computers to generate data for teaching statistics is described by Carmer and Cady (1969). Computer random number generators solve the frequently encountered problem of finding suitable examples of data for classroom discussion, homework problems, and laboratory assignments by generating sets of data according to a model whose structure and parameters are controlled by the teacher. If access to a random number generator is provided in a statistical analysis system, then a single computer program can be used to generate the sample data, perform computations, and display the data and results.

The use of SAS for the generation of data having univariate distributions is described elsewhere (SAS Communications, 1974). Because of the need for generating data with a more complex structure, we implemented a SAS procedure called RNG (Random Normal Generator) to illustrate multivariate statistical methods.

THE RNG PROCEDURE

The Algorithm

The RNG procedure uses the McGill University "Super-Duper" random number generator package to generate a vector of independent standard normal deviates (Marsaglia, et al., 1974). Super-Duper uses a combination of a shift register generator and a multiplicative congruential generator to produce a sequence of standard uniform random numbers. Because of the combination, the sequence has a long period (10^{18}). Also, the package uses a composition technique to generate standard normal deviates and thus offers exceptional speed.

RNG uses the technique of Scheuer and Stoller (1962) to translate a vector of independent standard normal deviates into a vector having an arbitrary covariance matrix, C say. The technique is an application of the following theorem.

Let \( Y' = (Y_1, Y_2, \ldots, Y_p) \) be distributed \( N(0, I_p) \) where \( I_p \) is the identity matrix of size \( p \). Then \( X = CY \) is distributed \( N(0, C'C) \) where \( C \) is an \( n \times p \) matrix of fixed elements.

Thus we need the decomposition \( E = C'C \). The matrix \( C \) is readily determined and unique if \( E \) is chosen to be upper triangular. If \( E \) is not positive definite, an error message is printed by the procedure.

The Control Cards

The RNG procedure produces as output a SAS dataset whose observations have values that are normally distributed. The procedure is called by a SAS statement of the form:

```
PORD RNG parameters;
```

The parameters allow the user to initialize the random number generator, to specify the number of decimal digits in the output dataset, and to rename certain variables.

A GENERATE procedure information statement is required. It specifies the names of the variables in the vector being generated. An example is:

```
GENERATE HEIGHT WEIGHT;
```

Optionally, a GROUPS, a SAMPLES, a MEANS, a VARIANCES, and a REPEAT procedure information statement can be used. The GROUPS statement specifies that the observations being generated are to be considered as being divided into two or more groups. The GROUPS statement provides the names for these groups. Different groups of observations may have different mean vectors, different covariance matrices, or different sample sizes. An example is:

```
GROUPS MALE FEMALE;
```

If the GROUPS statement is used, the output data set has a SAS character variable named GROUP whose values are one of the group names specified on the GROUPS statement.

The MEANS statement specifies the mean vector for the observations being generated or the group mean vectors. If a GROUPS statement is used, an example is:

```
MEANS MALE 70 150 FEMALE 65 105;
```

If the MEANS statement is not used, or if a specification does not appear for all groups, a zero mean vector is assumed for those groups not specified.
The VARIANCES statement specifies the covariance matrix for the observations being generated or the group covariance matrices, if a GROUPS statement is used. An example is:

```
VARIANCES 16 32 81;
```

where the variance of HEIGHT is 16, the covariance between HEIGHT and WEIGHT is 32 and the variance of WEIGHT is 81. Note the specification on the VARIANCES statement is only the upper triangular part of the covariance matrix. If the VARIANCES statement is not used, or if a specification does not appear for all groups, an identity matrix is assumed for those groups not specified.

The SAMPLES statement specifies the number of observations to be generated. If a GROUPS statement has been used, the SAMPLES statement may be used to specify different sample sizes. If the SAMPLES statement is not used, or if a specification does not appear for all groups, a value of 10 is assumed for those groups not specified. If a specification appears on a MEANS, VARIANCES, or SAMPLES statement without a group name and a GROUPS statement has been used, the specification applies to all the groups.

The REPEAT statement specifies that the generation of normal observations as specified on the GENERATE, SAMPLES, MEANS, VARIANCES, and/or GROUPS statements is to be repeated the specified number of times. An example is:

```
REPEAT 20;
```

If the REPEAT statement is used, the output data set has a SAS numeric variable named REPEAT whose values are ascending integers. All observations produced during the same repetition have the same value of REPEAT.

### Teaching and Research Applications

RNG is useful for generating multivariate normal data based on mean vectors and covariance matrices determined from real data or from theoretical considerations. The former approach offers a more realistic setting for the student whereas the latter is better for demonstrating certain conceptual points.

#### Real Data Examples

##### A Principal Components Example

The EPA monitored heavy metal concentrations in clams off the coast of Delaware. The data consisted here consists of 111 observations measured as concentrations of the following heavy metals: aluminum, cadmium, cobalt, nickel, vanadium, and zinc. The mean vector and the covariance matrix were computed using the LOG10 transformation of the observations and were used as input to RNG. The object of this exercise was to generate a data set (100 observations) for each student in a class in multivariate statistics in order to illustrate the sampling variation of the eigenvalues and eigenvectors of a correlation matrix (i.e., principal components analysis).

Based upon 18 student (i.e., REPEAT had a specification of 10), the estimated mean of the sampling distribution of the largest eigenvalue was 2.527 and the estimated standard deviation was 0.115. The asymptotic distributions of the eigenvalues and eigenvectors are known (Morrison; 1976). Thus the estimates of the parameters in the various sampling distributions can be compared to the known parameters of the asymptotic distributions. In this case, the theoretical mean and standard deviation were found to be 2.239 and 0.318, respectively, by using PROC MATRIX. The divergence of the estimated standard deviation from the hypothetical is unusual since the estimates are based on only 18 values. Other statistical quantities are not considered here due to space limitations.

This approach did permit the instructor to have different data sets for each student to analyze. Not only was a printed and punched copy of the data available to each student, but also a "solutions manual" was available for the instructor. The sampling distributions of the various statistics were then examined by compiling results from the "solutions manual".

#### A Multivariate Linear Model Example

This example is based on the TALENT data as given by Cooley and Lohnes (1971). Both a MANOVA and a multivariate regression were illustrated. The qualitative variable consisted of three categories of students' plans to attend college. The quantitative variables consisted of scores for Information I, Information II, English, Reading, Creativity, Mechanical, Abstract, and Mathematics tests.

Three groups of 25 observations each were generated for each student. The mean vectors and the common covariance matrix were computed using PROC MEANS and PROC CORR. These matrices as well as the control cards for RNG and for the procedures for the "solutions manual" were stored on a temporary file. This allows the instructor to obtain everything in one run and to avoid inputting the mean vectors and covariance matrices.

The MANOVA analysis consisted of the following SAS statements:

```
PROC GLM;
BY NAME;
CLASS PLANCOLL;
MODEL ENGLISH READING MECHAN ABSTRACT MATH = PLANCOLL/SS4;
MANOVA H=PLANCOLL/PRINTE PRINTH;
```

NAME contains the names of students in the multivariate class.
The values of the likelihood ratio test statistic for each student were:

0.609, 0.654, 0.464, 0.621, 0.630, 0.639, 0.511, 0.705, 0.689, 0.595, 0.686, 0.614, 0.691, 0.630, 0.647, 0.621, 0.650, 0.630, 0.577, 0.540, 0.591, and 0.733.

Since the critical value is 0.48, none of the students would reject the null hypothesis. However, the variation of the test statistics is clearly evident.

The SAS statements for the multivariate regression were:

PROC GLM;
BY NAME;
MODEL INPUT INPUT2 = ENGLISH READING CREAT MECHEAN ABSTRACT MATH/SS4;
MANOVA R=ENGLISH READING CREAT MECHEAN ABSTRACT MATH/PRINT PVALUE;

All of these tests are based on a single eigenvalue since the ranks of the hypothesis matrices are 1. The results in Table 1 summarize the conclusions of the students.

Table 1. Conclusions of the multivariate tests for each variable in the multivariate regression

<table>
<thead>
<tr>
<th>Variable</th>
<th>p&gt;.10</th>
<th>p&lt;.10</th>
<th>p&lt;.05</th>
<th>p&lt;.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENGLISH</td>
<td>14</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>READING</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>CREAT</td>
<td>14</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>MECHEAN</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>17</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MATH</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
</tr>
</tbody>
</table>

* This indicates that 14 of the 18 students failed to reject $R_0$ at the 0.10 level of significance.

Thus ENGLISH, CREAT, and ABSTRACT are generally not significant whereas READING, MECHEAN, and MATH are often significant at the 0.01 probability level.

These examples illustrate the sampling variabilities of various estimates and test statistics using real situations. However, many statistical concepts are more easily seen by controlling the values of the mean vectors and the covariance matrices.

Controlled Examples

Controlled parametric specifications are useful for studying power and the effects of departures from assumptions. In particular, the effects of unequal covariance matrices, of contamination, and of nonnormality on the general linear model are of interest.

Power

The idea of power was illustrated by the following SAS statements:

PROC RNG;
GENERATE XI-X3;
GROUPS A B;
MEANS 1 1.1 1;
VARIANCES 1 .6 1 .6 1;
REPEAT 25;
PROC GLM; BY REPEAT;
MODEL XI-X3 = GROUP;
MANOVA H = GROUP;

The hypothesis of the equality of means is false since $\mu_A=[0,0,0]$ and $\mu_B=[1,1,1]$ where $\mu_A$ is the mean vector of group A and $\mu_B$ is the mean vector of group B. The estimate of the power for this alternative based on 28 samples of 10 observations each for each group is 0.607.

The power function of Hotelling’s two sample $T^2$ can be estimated for a fixed $\lambda$ (the common covariance matrix) by varying $\mu_A$ and $\mu_B$. This results since the power function depends on $(\mu_A - \mu_B)^T\Sigma^{-1}(\mu_A - \mu_B)$. Power estimates for more complex linear models are also easily attainable by using PROC RNG.

Unequal Covariance Matrices

The effects of unequal covariance matrices could be illustrated by the following:

PROC RNG;
GENERATE XI-X3;
GROUPS A B;
MEANS 1 4 3;
VARIANCES 1 .6 1 .8 1 .3 1 .3 1 ;
SAMPLES 25;
PROC DISCRIM POOL=TEST WORK POORR LIST;
CLASS GROUP;
VARIABLES XI-X3;

For this example, the Chi-square homogeneity test rejects the hypothesis at a 0.001 level of significance. The power of this test can be studied by using the REPEAT option.

Contamination

The effects of contamination could be illustrated by the following:

PROC RNG OUT = DATA1;
GENERATE XI-X3;
MEANS 1 4 3;
SAMPLES 100;
PROC RNG OUT = DATA2;
GENERATE XI-X3;
MEANS 2 7 8;
SAMPLES 5;
DATA DATA3;
SET DATA1 DATA2;
In this case DATA1 contains "outliers" from DATA2. This would permit the instructor to illustrate the sensitivity of various estimates and test statistics to outlying observations.

**Nonnormality**

The effects of nonnormality could be assessed by the following:

```
PROC RKG OUT = DATA1;
GENERATE X1-X3;
SAMPLES 50;
DATA DATA2;
SET DATA1;
X2 = X2*X2;
X3 = EXP(X3);
PROC KSLTEST;
PROC PROBPLOT FULLNORM;
```

In an analysis, X1 was consistent with the normality assumption whereas X2 and X3 diverged in the expected way. PROC PROBPLOT is a procedure which is local to the West Virginia Network for Educational Telecomputing.

**DISCUSSION**

The preceding examples illustrate the usefulness of RNG to the instructor of statistics and to the researcher. Most concepts, depending on the normality of data, are easily demonstrated by using RNG in conjunction with the other capabilities of SAS.

This procedure should prove to be very beneficial to the instructor of multivariate statistics. The problem of finding suitable data for analysis by the students and of studying the complex sampling distributions of various statistics are to a large extent overcome.

**REFERENCES**


