USING MATHEMATICA® FOR MULTI-STAGE SELECTION AND OTHER MULTIVARIATE NORMAL CALCULATIONS

R.K. Shepherd and S.V. Murray
Central Queensland University, Rockhampton, QLD 4702

SUMMARY
The paper discusses the use of the technical computing package Mathematica® for performing calculations on multivariate normal distributions. An example is given of calculating response to selection using independent culling levels which requires the calculation of the mean and volume of a truncated trivariate normal distribution. The power and flexibility of using Mathematica® for optimising multi-stage selection is demonstrated and discussed.

Keywords: truncation selection, multivariate normal, multi-stage selection, Mathematica®

INTRODUCTION
Much of the theory of livestock breeding compares the response to selection using different methods of selecting parents. For traits determined by many genes of small effect, all acting in an additive manner (i.e. the infinitesimal model), the selection response is calculated using a normal distribution. In many situations the calculations need to be done using a multivariate normal distribution. For example, multi-stage selection involves selection (or culling) animals at various stages, often because some measurements are very expensive to collect on all animals or because information for calculating breeding values accrues progressively over an animal’s life. Calculating the response to multi-stage selection has often ignored the fact that the population is no longer normally distributed after the first stage of selection (Cotterill and James, 1981). This is because the numerical calculation of the mean and volume of a truncated multivariate normal distribution is computationally difficult. For example, Wade and James (1996) stated in their paper on optimising two stage selection that “The most difficult procedure in the algorithm was assessing the volume of the twice-truncated bivariate normal distribution”.

The difficulty in making multivariate normal calculations has reduced with the availability of technical computing packages like Mathematica® (Wolfram, 1999). The power of Mathematica® comes from it’s ability to choose the best algorithm for a problem and apply it adaptively, which, when combined with it’s symbolic computational power, programming capability and graphical options, result in an intelligent and flexible approach to technical computation. This paper will demonstrate some of these options available in Mathematica®.
MULTIVARIATE NORMAL CALCULATIONS IN MATHEMATICA®
Calculations involving the multivariate normal distribution are available in Mathematica® through a standard add-on package for Statistics called ‘MultinormalDistribution’. Table 1 lists a few of the functions which are available in this standard statistics package. Examples will be given to illustrate the use of these functions for multivariate normal calculations in animal breeding. For further information about Mathematica®, an excellent reference is The Mathematica® Book (Wolfram, 1999) which is available not only in print but also online at http://www.wolfram.com/.

Table 1. Some functions available in the standard add-on package ‘MultinormalDistribution’

<table>
<thead>
<tr>
<th>Syntax of Mathematica function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>dist = MultivariateNormalDistribution[μ, S]</td>
<td>specifies dist is a multivariate normal distribution with mean μ and covariance S</td>
</tr>
<tr>
<td>PDF[dist, z]</td>
<td>probability density function for dist at z, where z is a scalar or vector depending on dist</td>
</tr>
<tr>
<td>CDF[dist, z]</td>
<td>cumulative distribution function of dist at z ie. cumulative probability from −∞ to z</td>
</tr>
<tr>
<td>NIntegrate[f[x, y], {x, a, b}, {y, c, d}]</td>
<td>numerically integrates f(x, y) for a&lt;x&lt;b and c&lt;y&lt;d, ie. calculates ∫∫ dy dx f(x, y)</td>
</tr>
</tbody>
</table>

EXAMPLE 1. MEAN AND VOLUME OF A TRUNCATED TRIVARIATE NORMAL
This example is from Saxton (1982) and involves two stage selection using different traits at each stage (ie. independent culling). Assume the traits have zero means, unit phenotypic variances and a phenotypic correlation of 0.5. The heritability (h²) is 0.6 and 0.3 for trait one (x) and trait two (y) respectively. The breeding objective G is given by G = Gₓ + Gᵧ, where Gₓ and Gᵧ are the breeding values for x and y respectively. It can be shown that σₓ² = hₓ² + hᵧ² + 2rₓhₓhᵧ , σᵧ² = hᵧ² + rₓhₓhᵧ , and σₓᵧ = hₓhᵧ + rₓhₓhᵧ , where σₓ², σᵧ², and σₓᵧ are the variance of G , the covariance between x and G , and the covariance between y and G respectively. Using these equations and a genetic correlation rₓ of 0.4, it can be shown that the vector (x, y, G) has a trivariate normal distribution with mean μ = (0, 0, 0) and covariance matrix S as given in Table 2.

Saxton (1982) shows that choosing only animals with x > 1.2786 and y > –1.3958 will select 10% of the population and that the mean G of the animals selected will be 1.351. In mathematical notation,
Saxton shows that the proportion 0.1 is given by

\[ 0.1 = \int_{-\infty}^{\infty} \int_{-1.3958}^{1.2786} \int_{-1.3958}^{1.2786} f(x, y, G) \, dx \, dy \, dG, \]

where \( f(x, y, G) \) is the trivariate normal density function. The mean \( G \) of the animals selected, is then calculated by the following integral;

\[ 1.351 = \int_{-\infty}^{\infty} \int_{-1.3958}^{1.2786} G \times f(x, y, G) \, dx \, dy \, dG / 0.1. \]

The ease of performing these mathematical operations in Mathematica® is illustrated in Table 2.

To calculate the proportion \( prop = 0.1 \) in Table 2, the Mathematica® function \( CDF[dist, z] \) is used. This function calculates the proportion from \(-\infty\) to \( z \). However the proportion needed is from \( z \) to \( \infty \) (for truncation selection) which, by symmetry of the standardised multivariate normal distribution, is given by \( CDF[dist, -z] \). Also note that instead of the function \( CDF[mvn3, \{-xCut, -yCut, \infty\}] \), the function \( NIntegrate[f[x, y, G], \{x, xCut, \infty\}, \{y, yCut, \infty\}, \{G, -\infty, \infty\}] \) could have been used.

Table 2. Mathematica® code (italics), and output (bolded), to calculate the proportion selected and mean \( G \), for the independent culling described in Example 1.

```
<< Statistics`MultinormalDistribution`
μ = {0, 0, 0}; xCut = 1.2786; yCut = -1.3958;
S = {{1, 0.5, 0.7697}, {0.5, 1, 0.4697}, {0.7697, 0.4697, 1.2394}};
mvn3 = MultinormalDistribution[μ, S];
prop = CDF[mvn3, {-xCut, -yCut, \infty}]; Print[“Selected proportion = “, prop]
Selected proportion = 0.100002
f[x_, y_, G_] = PDF[mvn3, {x, y, G}];
SumG = NIntegrate[G*f[x, y, G], {x, xCut, \infty}, {y, yCut, \infty}, {G, -\infty, \infty}];
Print[“Mean G of the animals selected = “, SumG / prop]
Mean G of the animals selected = 1.35111
```

Example 1 assumes the truncation points on each axis are known. In multistage selection we wish to determine the truncation points (or equivalently the proportion selected at each stage) in order to maximise the breeding objective over all stages of selection. This is the topic of Example 2.

**EXAMPLE 2. OPTIMISING MULTISTAGE SELECTION**

Saxton (1982) gives results for optimising two-stage selection when selection is on the basis of different traits at each stage. The parameters for the two traits are identical to those in Example 1 except for the genetic correlation \( r_A \) which is varied. Saxton(1982) determines truncation points for the two traits such that choosing all animals with \( x > t_1 \) and \( y > t_2 \) results in the 10% of the
population with the largest mean value for the breeding objective \((G)\). In mathematical notation, Saxton determines the values of \(t_1\) and \(t_2\) which:

\[
\text{maximise } \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G \times f(x, y, G) \, dx \, dy \, dG / \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y, G) \, dx \, dy \, dG \quad \text{ (ie. maximise the mean } G) \\
\text{subject to satisfying the constraint } \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y, G) \, dx \, dy \, dG = 0.1 \\
\text{(ie. exactly 10% of all animals are selected). This problem was coded up in one page of Mathematica\textsuperscript{®} instructions using it’s programming language. The results are presented in Table 3 along with those reported by Saxton (1982). For each genetic correlation the maximum mean } G \text{ found was similar to that found by Saxton (1982), with the only discrepancy being for } r_A = -0.05 \text{, which is most probably due to a typing error in Saxton’s paper. Some minor discrepancies occurred in the optimum truncation points due to the flatness of the optimum peak and the greater accuracy available in Mathematica.}

Table 3. Truncation points for trait \(x\) \((t_1)\) and \(y\) \((t_2)\) in Example 2 which maximise the mean \(G\) for genetic correlation \((r_A)\) calculated using Mathematica\textsuperscript{®} and compared to Saxton (1982).

<table>
<thead>
<tr>
<th>(r_A)</th>
<th>(t_1)</th>
<th>(t_2)</th>
<th>(\text{Mean } G)</th>
<th>(t_1)</th>
<th>(t_2)</th>
<th>(\text{Mean } G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4</td>
<td>1.2770</td>
<td>-1.2631</td>
<td>1.351</td>
<td>1.2786</td>
<td>-1.3958</td>
<td>1.351</td>
</tr>
<tr>
<td>-0.05</td>
<td>1.2816</td>
<td>-3.4897</td>
<td>1.016</td>
<td>1.2816</td>
<td>-4.0000</td>
<td>1.036</td>
</tr>
<tr>
<td>-0.4</td>
<td>1.2816</td>
<td>-4.3887</td>
<td>0.756</td>
<td>1.2816</td>
<td>-4.0000</td>
<td>0.755</td>
</tr>
</tbody>
</table>

**DISCUSSION**

Most literature on optimising multi-stage selection, which has not ignored the multivariate normal nature of the problem, has only been applied to scenarios involving two stages of selection. This is mainly due to the difficulties in calculating volumes under the normal density in three or more dimensions. However although not demonstrated in this paper we have investigated higher dimensional normal distributions using Mathematica\textsuperscript{®} functions. We were able to instantly (on a 1.8GHz Pentium 4 PC with 512Mb memory) calculate the volume under a truncated 10 dimensional standard multivariate normal distribution using the Mathematica\textsuperscript{®} function \(\text{CDF}\). However to use \(\text{CDF}\) the correlation matrix has to have a specific structure in order that its inbuilt efficient algorithm can be used. For totally unstructured correlation matrices the Mathematica\textsuperscript{®} function \(\text{NIntegrate}\) can be used to calculate means and volumes for truncated standardised multidimensional normal
distributions. However, \textit{NIntegrate} slows down considerably when the number of dimensions exceeds four. There are options in Mathematica\textsuperscript{®} to speed this up.

This paper has demonstrated the power and ease of performing calculations on multivariate normal distributions in Mathematica\textsuperscript{®}. The demonstration has been specific to two-stage selection but it is applicable to any calculations involving multivariate normal distributions. In particular, deviations from normality due to selection or non-random mating could be investigated with this software. For example, it may be possible to examine precisely the combined effects of selection and assortative mating over many generations by extending the theory developed by Smith and Hammond (1987). These types of investigations were previously not possible, but now are, due to the power and flexibility available in technical computing packages like Mathematica\textsuperscript{®}.

\textbf{REFERENCES}