For the last five columns we've looked at the problem of characterizing multivariate data. An implicit assumption that runs across principal components analysis, factor analysis, principal coordinates analysis, correspondence analysis and partial least-squares analysis is that the objects included in the dataset represent independent and randomly selected samples drawn from a population of interest. So long as we were asking questions about the particular assemblage of data (e.g., the trilobite data we've been using as a running example), the results of the analyses we have obtained to date are perfectly valid if largely indicative given the relatively small sample size. For our illustrative purposes these 20 genera were the population and this is how we've been discussing them; as if no other types of trilobites exist. But of course, there are other types of trilobites. The time has come to acknowledge this fact and explore the types of analyses we might apply to datasets that exhibit various types of internal structure.

The simplest type of structure is that of subgroups existing within the dataset. Taxonomic datasets are often composed not of a single representative of each group (e.g., genus or species) or multiple representatives of a single group, but multiple representatives of a few well-defined groups. Often in systematics and (palaeo)ecology our problem is not so much one of trying to explain the structure of relations between measurements or observations collected from single groups, as trying to use a common set of measurements or observations to characterize groups of taxa, guilds, etc. Indeed, this is the standard problem of systematics: how many groups are there and how best to distinguish them. Of course we'll need to state these questions a bit more precisely in order to answer them quantitatively.

As usual, I find the best way to discuss the issues involved in group evaluation and characterization is through an example dataset. Our trilobite data are not adequate for this purpose as they don't lend themselves to being collected into groups that make much sense. Instead, we'll reference our discussion to a classic dataset that R. A. Fisher used to explain the concepts behind a set of methods that have come to be known as discriminant analysis (Fisher 1936). Fisher did the obvious when he became interested in the 'groups' question, he went out and obtained some measurements from different groups: in his case four simple measurements on three *Iris* species. Actually, the 'Fisher' *Iris* data weren't collected by Fisher, but rather by *Iris* researcher Edgar Anderson (1935). Regardless, ever since Fisher's first article on these flowers statisticians, researchers and teachers have been using the Fisher *Iris* data as a reference dataset for developing, testing and illustrating discriminant analysis methods. The full dataset consists of 50 sets of measurements for four variables collected from each species. However, there's no need to pile up the sample numbers for our simple purposes. The fist ten sets of measurements for each species will suffice. These are reproduced in Table 1.

Figure 1. Photographs of the three *Iris* species used by Fisher (1936) to illustrate the properties of discriminant analysis. Images courtesy of the Species *Iris* Group of North America (http://www.badbear.com/signa/signa.pl?Introduction).

Table 1. First ten specimens from each species included in Fisher (1936) *Iris* data.

	Iris setosa				*Iris versicolor*			
	Petal		Sepal		Petal		Sepal	
	Length	Width	Length	Width	Length	Width	Length	Width
1	5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4
2	4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5
3	4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5
4	4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3
5	5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5
6	5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3
7	4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6
8	5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0
9	4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3
10	4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4
Σ	48.6	33.1	14.5	2.2	61.0	28.7	43.7	13.8
Min.	4.4	2.9	1.3	0.1	4.9	2.3	3.3	1.0
Max.	5.4	3.9	1.7	0.4	7.0	3.3	4.9	1.6
Mean	4.9	3.3	1.5	0.2	6.1	2.9	4.4	1.4
Median	4.9	3.3	1.4	0.2	6.4	2.9	4.6	1.4
Variance	0.1	0.1	0.0	0.0	0.5	0.1	0.2	0.0
S. Dev.	0.3	0.3	0.1	0.1	0.7	0.3	0.5	0.2

	Iris virginica			
	Petal		Sepal	
	Length	Width	Length	Width
1	6.3	3.3	6.0	2.5
2	5.8	2.7	5.1	1.9
3	7.1	3.0	5.9	2.1
4	6.3	2.9	5.6	1.8
5	6.5	3.0	5.8	2.2
6	7.6	3.0	6.6	2.1
7	4.9	2.5	4.5	1.7
8	7.3	2.9	6.3	1.8
9	6.7	2.5	5.8	1.8
10	7.2	3.6	6.1	2.5
Σ	65.7	29.4	57.7	20.4
Min.	4.9	2.5	4.5	1.7
Max.	7.6	3.6	6.6	2.5
Mean	6.6	2.9	5.8	2.0
Median	6.6	3.0	5.9	2.0
Variance	0.6	0.1	0.4	0.1
S. Dev.	0.8	0.3	0.6	0.3

The basic problem these data present can be summarized by plotting all combinations of variables in the form of a matrix of scatterplots (Fig. 2).

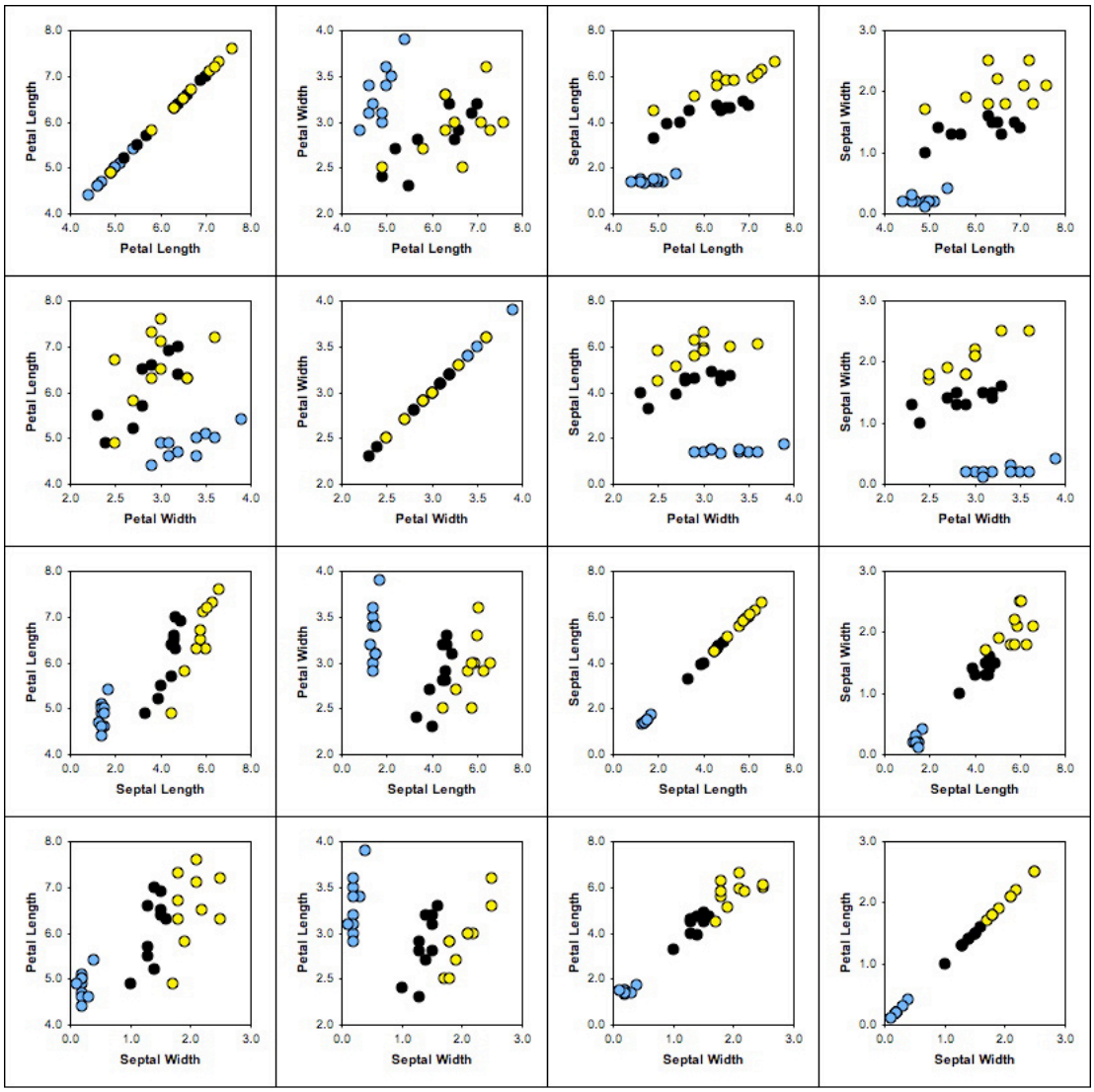

Figure 2. Crosstabulation diagram for Fisher *Iris* data. *I. setosa* (cyan), *I. versicolor* (black), *I. virginica* (yellow).

Given the bewildering variety of geometric relations between these three groups relative to these four variables what can we conclude regarding the distinctiveness of the groups? Moreover, if the groups are distinct can we use these data to construct a model of variation for each group that will allow us to assign unknown datasets to the correct group?

The first step in this process requires investigation of the structure of relations among groups. If all the groups have the same statistical structure our job is going to be much easier and more accurate. Of course, this begs the question of what 'same structure' means. Two factors are considered important, (1) the separation of group means relative to the variance of each group across all variables and (2) the pattern of between-variable covariance of each group. These factors are independent of one another insofar as the means may be distinct among groups whose covariance structure is identical and *vice versa*.

The standard test for assessing the significance of difference between multivariate means is an extension of the popular single variable, or univariate, Student's *t*-test; the Hotelling (1931) T^2 statistic. Derivation of the statistic is somewhat complex and need not concern us in detail (interested readers should consult Morrison, 2005). The overall form of the statistic, however, is important as we will see variations of it throughout this column and the next.

$$T^2 = n_1 n_2 (\bar{x}_1 - \bar{x}_2)' S_p^{-1} (\bar{x}_1 - \bar{x}_2) \Big/ (n_1 + n_2) \qquad (10.1)$$

I've deviated a bit from the usual T^2 formula in order to make the relations more explicit and represent the test as a comparison between two samples rather than between a sample and a population. The $(\bar{x}_1 - \bar{x}_2)$ term is simply the difference between the means of two groups, 1 and 2. Because these means involve all measured variables, each contains (in our case) four terms, one for each variable. By mathematical convention these differences are represented as a matrix of one column and whose number of rows is equivalent to the number of variables. These difference matrices can also be regarded as a set of vectors whose directions and magnitudes express inter-group similarities and differences. The difference matrices/vectors for the *Iris* data are shown in Table 2.

Table 2. Difference matrices/vectors for the *Iris* data.

	I. setosa vs. *I. versicolor*	*I. setosa* vs. *I. virginica*	*I. versicolor* vs. *I. virginica*
Petal Length	-1.24	-1.71	-0.47
Petal Width	-6.18	0.37	-0.07
Sepal Length	-2.92	-4.32	-1.40
Sepal Width	-1.16	-1.82	-0.66

Inspection of this table suggests the mean values for *I. setosa* are substantially smaller than those of *I. versicolor* and *I. virginica*. Note this agrees with both Table 1 and Figure 1.

The $(\bar{x}_1 - \bar{x}_2)'$ term represents the transposed form of the difference matrices. That is, the transpose of these matrices has one row and four columns of figures. A matrix (X) pre-multiplied by its transpose (X') yields the matrix of squares and cross-products; a standard statistical measure of covariation between sets of variables.

The S_p^{-1} term represents the inverse of the pooled variance-covariance matrix. The inverse of a matrix is used to perform the division operation in matrix algebra. Just as division of (say) 4 by 2 can be performed by taking the reciprocal of 2 (= 0.5) and multiplying that value by 4, one matrix can be divided by another by taking the inverse of the latter and post-multiplying it by the former. Because we are considering two samples in the *Iris* comparison we also need to generate an estimate of these samples' combined covariance structure. This is a simple operation that effectively determines an average of the two group (S_1 and S_2) covariance matrices weighted by the group sample sizes (n_1 and n_2). The following equation specifies this calculation.

$$S_p = \left[(n_1 - 1)S_1 + (n_2 - 1)S_2\right] / (n_1 + n_2 - 2) \qquad (10.2)$$

Because sample sizes for the *Iris* species groups are the same for each dataset the pooling calculation simplifies to determining the average of corresponding covariance matrix elements across the three datasets. Results of pooling the covariance matrices and taking their inverse are shown in the *PalaeoMath 101: Groups I* worksheet (see url below). Equation 10.1 represents the multivariate analogue of Student's *t*-test, in which the difference between the mean of a sample is compared to a reference value (theoretically the population mean, but often the mean of another sample) with the result being scaled by the sample size (n) and a measure of the samples' common variance structure.

One final small complication. Whereas the expected distribution of Student's *t*-values for samples of various sizes is well known the expected distribution of Hotelling's T^2 values, is more obscure. Fortunately, this is not a problem because the T^2 statistic can be transformed into an equivalent *F*-statistic using the following relation.

$$F = (n_1 - n_2 - m - 1)T^2 / (n_1 + n_2 - 2)m \qquad (10.3)$$

Here n_1 *and* n_2 is the number of specimens in the samples 1 and 2 respectively and m is the number of variables in the datasets. Of course, the *F*-test also requires specification of two

degrees of freedom (dof). For the Hotelling's T^2 conversion the numerator dof is the number of variables (*m*) and the denominator dof is the total number of specimens minus the number of variables in the sample, minus 1 (= $n_1 + n_2 - m - 1$). Applying these equations to the *Iris* data results in calculation of the following values.

Table 3. Results of Hotelling's T^2 test of comparisons between species-group means.

	I. setosa - *I. versicolor*	*I. setosa* - *I. virginica*	*I. versicolor* - *I. virginica*
T^2	4864.41	1956.43	205,56
F	1148.54	461.94	48.54
Prob.	2.87×10^{-18}	1.66×10^{-15}	2.08×10^{-8}

Obviously the means are rather different from one another, even though the sample sizes are quite small, even for the superficially similar species *I. versicolor* and *I virginica*. This test confirms the idea that the overall character of the groups, as represented by these four variables, is decidedly different. However, it does not assess whether the groups have a similar covariance structure, whether the groups are best characterized by mutually exclusive or overlapping distributions, which variables are best at characterizing group identity, or whether unknown observations can be assigned to these groups with a high degree of accuracy. To answer these questions we need to perform additional analyses.

Because Hotelling's T^2 test assumes a common covariance structure for all samples we need to test that next, if only to confirm the previous result. There are a large number of statistical tests that have been proposed for this purpose, far more than are usually described in multivariate analysis textbooks much less a brief column like this. Of these the one I prefer is the likelihood ratio test (Manley 1994) because it is (1) powerful yet relatively easy to calculate, (2) uses some of the same terms we'll meet later in our discussion of canonical variates analysis, and (3) can be used to test either the equality of multivariate means or dispersion structure.[1]

The equation of the likelihood ratio test is as follows.

$$\phi = [n_t - 1 - 0.5(m + k)]\ln[|T|/|W|] \tag{10.4}$$

In this expression n_t represents the total number of specimens across all groups ($n_1 + n_2$), *m* (as before) represents the number of variables, and *k* represents the number of groups. Also *T* and *W* refer to two summary matrices that get to the heart of discriminant analysis. Matrix *T* represents the total sums of squares and cross products matrix and has the following form.

$$t_{r,c} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{i,r,j} - \bar{x}_r)(x_{i,c,j} - \bar{x}_c) \tag{10.5}$$

In this expression *r* and *c* refer to the rows and columns of the *T* matrix (any cell of which is occupied by a value *t*). The really important parts of this formula, though, are the variables $\bar{x}_r$ and $\bar{x}_c$ which are the grand means for the entire, combined dataset. In geometric terms the grand mean is the centre of the pooled sample of all measurements. Matrix *T*, then summarizes the dispersion of the total dataset about this group-independent, fixed reference.

[1] While we could have used the likelihood ratio test to perform the analysis we undertook using Hotelling's T^2 the null hypothesis would have involved testing the means for all three species-groups simultaneously, not in a pair-wise manner. For exploratory analysis a pair-wise strategy often yields more information.

Similarly, the W matrix summarizes the within-groups sums of squares and cross-products matrix and has the corresponding form:

$$w_{r,c} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{i,r,j} - \bar{x}_{jr})(x_{i,c,j} - \bar{x}_{jc}) \qquad (10.6)$$

Once again, *r* and *c* refer to the rows and columns of the *W* matrix (any cell of which is occupied by a value *w*). Now the variables $\bar{x}_{jr}$ and $\bar{x}_{jc}$ refer to the analogous group-specific means. In geometric terms the group mean is the centre of the cloud of points representing each group in Figure 1. Matrix *W*, then, summarizes the dispersion of each dataset relative to its own group-specific reference.

To get a handle on this statistic, in your mind's eye think about three clouds of points. The within-groups means are the centres of each individual cloud and the total groups mean is the center of all clouds taken together. If the position and orientation of the clouds are just about the same the ratio *T/W* is going to be a relatively small number. If the position and orientation of the clouds is radically different *T* will be much larger than *W* and the ratio will be large. The rest of the terms in equation 10.3 have to do with scaling the ratio for the overall dimensionality of the problem, both in terms of numbers of variables and specimens.

Notice the *T* and *W* symbols are enclosed by vertical lines in equation 10.4. Those are the symbols for the determinant of the *T* and *W* matrices. Most textbooks define the determinant of a matrix as the sum of all terms in the matrix (*n*!) taken in a highly peculiar order. Those discussions then usually go on for pages about the order in which the terms are taken—the algorithms developed to facilitate this calculation—and the implications of particular results (e.g., symmetric matrixes have positive determinants, a value of 0.0 means the matrix is singular, which, in turn, means it has no inverse). What they never seem to get around to telling you is that the determinant is nothing more than the 'volume' of the matrix, albeit a highly peculiar volume (see http://en.wikipedia.org/wiki/Determinant). If the determinants of the *T* and *W* matrices are similar, the structure of their covariance relations will (likely) be similar; if radically different the structure of their covariance matrices will (likely) be different. The ϕ-statistic is distributed according to χ^2 distribution with $m(k\text{-}1)$ degrees of freedom.

One last little bit about the likelihood ratio test. If you are going to use it to test the hypothesis of whether the group mean vectors are equivalent you use the raw data. If you are going to use it to test the equivalence of the group dispersion structures you must first convert your data to their median deviate form and then apply equations 10.4, 10.5 and 10.6. Since we've already tested the mean vectors using Hotelling's T^2 the *PalaeoMath 101:Groups I* worksheet illustrates the dispersion test (see Manly 1994 for an example of an application to mean-vector analysis on a similar simple dataset). Based on my calculations for these *Iris* data ϕ = 4.28 which has an associated χ^2 probability of 0.83. Since this probability value is much greater than the traditional 0.05 cut-off the *Iris* data fail the test and the null hypothesis of no difference in the dispersion (= covariance) structure among species-group datasets is accepted.

To this point in our analysis we've been entirely concerned with questions about whether it is appropriate for us to proceed with a full-blown multivariate discriminant analysis. Those results have told us there are significant differences between the means of all groups but no significant differences in the structure of geometric relations between variables across the same groups. This is the ideal situation; hence the widespread use of the *Iris* data for illustrating discriminant analysis. If your data don't match up to these fairly exacting standards don't throw your hands up in horror. It's not the end of the world. You'll just have to be extra cautious in interpreting results of the procedures I'll describe next and in the subsequent column.

Before we tackle the final analysis for this column and answer the question of how distinctive our species-groups are, though, let's stop for a moment and consider what we mean when we say 'These things form a group.' In taxonomy, ecology, phylogeny, biogeography, what have

you, similarity is judged by the objects belonging to a group all sharing some group-defining feature. It really doesn't matter what the feature is. It might be a distinctive structure, a preference for a certain habitat, a mode of locomotion, a behaviour, a colour, sound, or even a smell, etc. Whatever 'it' is, members of the group share it, non-members don't. Since this 'it' is a property of organisms the natural way for a mathematician/statistician to think about 'it' is in terms of a distance. If we represent specimens by some set of measured variables, or even qualitative observations, those that belong to groups should be 'close' to other members of the same group and 'farther away' from members of different groups. Distance is the natural metric for assessing group membership problems.

We've discussed distances before. Euclidean distances play a large role in principle coordinates analysis and various forms of multidimensional scaling. Distances also play a large role in discriminant analysis problems because, like the *Q*-mode methods we described and discussed earlier, distances are conceptually bound up with the way we usually think about group membership. But just like variables distances have their problems.

Actually, distances have their problems mostly because there is no way to calculate them except through variables and, as we've seen repeatedly, variables have their problems. The most fundamental of these is that variables tend to exhibit complex patterns of covariation with one another. If we calculate a distance under the assumption that its constituent variables have nothing to do with one another, and it turns out those variables exhibit similar patterns of variation, the distances that describe both between-groups and within-groups proximity will be mis-represented. Thus, in Figure 2 our three *Iris* species-groups are all more-or-less distinct from one another on certain plots—especially *I. setosa* from *I. versicolor* and *I. virginica*—but much less so in others. These patterns are caused by inter-variable covariance relations. Unfortunately, there is no way to estimate the extent to which raw geometries such as those depicted in Figure 2 are biased by variable covariances without performing some fairly complex mathematics.

Just as in 'real-life', distance calculations involving groups are facilitated by defining reference points. We need to agree on a single reference definition for a group's location in the mathematical space formed by its variables. In terms of classical discriminant analysis this reference location is usually taken as the group's mean or centroid. At first this might seem an unusual choice. After all, the centroid is always embedded well within the group's distribution, not close to its margins. These margins provide the most intuitive definition of the limits of group membership. Nevertheless, the centroid is a much more stable point than any on the distribution's margins and has the advantage of being able to indicate likely group membership even in cases where the margins of different groups overlap.

As we have seen, the Euclidean distance is widely used as a basis matrix for multivariate procedures. This is fine when the Euclidean distance is coupled with an eigenanalysis or singular value decomposition because these procedures transform the variables used to calculate distances in a manner that corrects for inter-variable covariances. But what if we don't want to conduct a principal coordinates and correspondence analysis, perhaps because those techniques are formulated to operate on single samples and we have a dataset that contains representatives of multiple groups? Is there a distance metric we can use to cover this situation?

On first pass you might be tempted to standardize the variables in your dataset before you calculate the Euclidean distance. This renders the variance of all variables equal to 1.0 thereby ensuring equal weighing for all variables in the distance calculation.[2] If your variables are referenced to incompatible units (e.g., composed of variables measured in millimetres, degrees, areas, etc. all lumped together) this will be the only realistic option. However, equal weighting for all variables is, in most cases, as artificial as wildly differential weighting. What is needed is a distance metric that respects the structure of covariance relations between variables.

[2] Another issue with the Euclidean distance metric that concerns some is that variables with a high variance are differentially influential in determining the final distance value.

Prasanta Chandra Mahalanobis introduced a distance measure that does precisely this in 1936 and ever since the 'Mahalanobis distance' has gone on to become a staple similarity index in a wide variety of multivariate data analysis contexts. We've seen the general form of the Mahalanobis distance before.

$$D^2 = (x - \bar{x})' S_p^{-1} (x - \bar{x}) \qquad (10.7)$$

Note its similarity to Hotelling's T^2 (equation 10.1). Like the T^2-statistic, the Mahalanobis distance represents the square of the deviation of an observation from the mean scaled by the inverse of the covariance matrix. This means all information about inter-variable covariances or collections is taken into account in the final value. Like the T^2-statistic, if more than a single sample is being evaluated the Mahalanobis distance should be based on the pooled covariance matrix so the best possible estimate of the true covariance structure is used, provided the data meet the assumption of no significant differences in covariance structure. The Mahalanobis distance also conforms to the χ^2 distribution with k degrees of freedom; a feature that makes it very useful for making statistical association tests. Thus, an observation with a low Mahalanobis D^2 relative to the group centroid is likely to be a member of that group irrespective of the distribution of the data (recall the χ^2 test is non-parametric), whereas a specimen that exhibits a significantly high Mahalanobis D^2 relative to any (or all) groups in the sample is likely not a member of that group (or those groups).

In interpreting the Mahalanobis distance it is important to remember it is a dimensionless 'distance' and, and so not expected to conform to a Euclidean distance (which is a scaled distance) in terms of magnitude. Rather what is looked for is the relative size of the distance between an object and various group centroids (many discriminant analysis programmes simply assign objects to groups based on the magnitude of D^2) and, in terms of statistical testing, the relation between D^2 and the appropriate χ^2 critical value.

So, how do our *Iris* groups stack up with respect to the Mahalanobis distance? Table 3 shows results for fitting the data from each specimen in Table 1 to the three species-group centroids using the pooled sample covariance matrix (calculated using equation 10.2, see *PalaeoMath 101: Groups I* worksheet for computational details). Remember this fitting is done without an accompanying eigenanalysis to 'clean up' inter-variable covariances. The degree to which each species can be assigned to the correct species-group provides an indication of how distinctive the group data are from one another.

Table 4. Mahalanobis D^2 values for the fitting of all data used to the species-group *Iris* models to the respective group centroids. Bold type indicates group centroids with the lowest D^2 distance. The $\chi^2_{df=4,\ \alpha=0.5}$ critical value = 4.895.

	Data: *Iris setosa* Group: *Iris setosa*	Data: *Iris setosa* Group: *Iris versicolor*	Data: *Iris setosa* Group: *Iris virginica*
1	**1.725**	79.869	333.795
2	**3.383**	31.323	506.995
3	**0.343**	36.433	640.398
4	**1.913**	37.274	834.175
5	**2.414**	50.756	948.785
6	**3.954**	61.131	1472.194
7	**0.969**	29.465	1652.406
8	**0.289**	48.115	2215.220
9	**3.970**	44.140	2413.238
10	**0.899**	49.497	188.287

	Data: *Iris versicolor* Group: *Iris setosa*	Data: *Iris versicolor* Group: *Iris versicolor*	Data: *Iris versicolor* Group: *Iris virginica*
1	156.319	**3.875**	61.294

2	170.640	**1.221**	47.836
3	194.685	**1.560**	39.440
4	214.270	**5.286**	39.434
5	206.608	**3.017**	38.683
6	221.105	**7.730**	31.342
7	199.018	**3.359**	34.704
8	140.126	**6.600**	73.121
9	175.334	**1.323**	49.041
10	195.665	**2.655**	38.219

	Data: *Iris virginica* Group: *Iris setosa*	Data: *Iris virginica* Group: *Iris versicolor*	Data: *Iris virginica* Group: *Iris virginica*
1	484.706	79.869	**9.312**
2	360.103	31.323	**2.476**
3	380.165	36.433	**3.060**
4	369.228	37.274	**4.068**
5	421.434	50.756	**1.119**
6	451.259	61.131	**4.776**
7	337.297	29.465	**8.549**
8	403.068	48.115	**7.168**
9	406.199	44.140	**3.670**
10	400.530	49.497	**7.315**

As you can see, our results are encouraging. All data used in this analysis fit their appropriate model and only a few individuals exhibit distances to the nearest group centroid that lie outside the $\alpha = 0.05$ confidence interval as assessed by the χ^2 distribution.[3] This implies our species-group data are actually much more discrete than implied by Figure 2. In the next column we'll discuss strategies we can employ for producing an ordination plot that will provide a visual indication of the true distinctiveness of these data.

Norman MacLeod
Palaeontology Department, The Natural History Museum
N.MacLeod@nhm.ac.uk

REFERENCES

ANDERSON, E. 1935. The irises of the Graspe Peninsula. *Bulletin of the American Iris Society*, **59**, 2–5.

FISHER, R. A. 1936. The utilization of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179–188.

HOTELLING, H. 1931. The generalization of Student's ratio. *Annals of Mathematical Statistics*, **2**, 360–378.

MAHALANOBIS, P., C. 1936. On the generalized distance in statistics. *Proceedings of the National Academy of Science, India*, **12**, 49–55.

MANLEY, B. F. J. 1994. *Multivariate statistical methods: a primer*. Chapman & Hall, Bury, St. Edmonds, Suffolk, 215 pp.

MORRISON, D. F. 2005. *Multivariate statistical methods*. Duxbury Press, New York, 498pp.

Don't forget the *Palaeo-math 101* web page, now at a new home at:
http://www.palass.org/modules.php?name=palaeo_math&page=1

MacLEOD, N. 2007. Groups I. Palaeontological Association Newsletter, **64**, 35—45.

[3] Given the very small sample size used in our example some error in estimation of the group centroid—yielding a few high Mahalanobis distances—is an expected result.