Multivariate Analysis of Ecological Data

MICHAEL GREENACRE
Professor of Statistics at the Pompeu Fabra University in Barcelona, Spain

RAUL PRIMICERIO
Associate Professor of Ecology, Evolutionary Biology and Epidemiology
at the University of Tromsø, Norway

Chapter 17 Offprint

Inference in Multivariate Analysis

First published: December 2013
ISBN: 978-84-92937-50-9

Supporting websites:
www.fbbva.es
www.multivariatestatistics.org

Fundación BBVA

CHAPTER 17

Inference in Multivariate Analysis

We have presented multivariate analysis as the search for structure and relationships in a complex set of data comprising many sampling units and many variables. Groupings are observed in the data, similarities and differences reveal themselves in ordination maps, and the crucial question then arises: are these observed patterns just randomly occurring or are they a signal observed in the data that can be considered significant in the statistical sense? In this chapter we shall tackle this problem in two different ways: one is using bootstrapping, to assess the variability of patterns observed in the data, analogous to setting confidence intervals around an estimated statistic, and the other is using permutation tests in order to compute *p*-values associated with the testing of different hypotheses. We will illustrate these computational approaches to statistical inference in two different situations, where group differences or associations between variables are being assessed. Before tackling the multivariate context we shall treat more familiar univariate and bivariate examples in each respective situation.

Contents

Univariate test of group difference

One of the simplest statistical tests to perform is the two-group *t*-test of difference in means between two populations, based on a sample from each population. Taking our "bioenv" data set as an example, suppose we aggregate the samples corresponding to clay (C) and sand (S) sediment (labelled as group CS), to be compared with the gravel sediment sample (G). We want to perform a hypothesis test to compare the pollution values for the 22 sites of CS with the 8 sites of G. The mean pollution in each group is computed as:

$$\bar{x}_{CS} = 5.18 \qquad \bar{x}_{G} = 2.70$$

Performing the usual *t*-test and not worrying for the moment whether the assumptions of this test have been satisfied, we obtain a *t*-statistic (with 30 – 2 = 28 degrees of freedom) of 3.22 with associated *p*-value for a two-sided test of 0.0032. Thus we would conclude that there is truly a difference in pollution between clay/sand and gravel samples, with gravel samples having less pollution on average. The estimated difference in the mean pollution is 2.48 and a 95% confidence interval for the difference in the means is [0.90, 4.05]. Now this simplest form of the *t*-test assumes that the data are normally distributed and that the variances are equal in the two groups. An alternative form of the *t*-test, known as the *Welch test,* does not assume equal variances and obtains a *t*-statistic of 4.62, a lower *p*-value of 0.00008 and a much narrower 95% confidence interval of [1.38, 3.58]. If we examine the normality by making a normal quantile plot and using the *Shapiro-Wilks test*[1] in these quite small samples, there is no strong evidence that the data are not normal.

An alternative distribution-free approach to this test, which does not rely on the normality assumption, is to perform a *permutation test.* Under the hypothesis of no difference between the two groups it is assumed they come from one single distribution of pollution, so any observation could have been in the clay/sand group or the gravel group. So we randomly assign the 30 pollution observations to a sample consisting of 22 of the values, with the remaining 8 values in the other sample, and recompute the difference in the group means. The number of ways we can randomly separate the 30 values into two samples of 22 and 8, is:

$$\binom{30}{22} = \binom{30}{8} = 5{,}852{,}925$$

that is, almost 6 million ways. In such a situation we do this random allocation a large number of times, typically 9,999 times, plus the actual observed samples, giving 10,000 permutations in total. The distribution of these 10,000 values, called the *permutation distribution,* is given in Exhibit 17.1 – it is the distribution of the difference in means under the null hypothesis of no difference. To obtain a *p*-value we see where our observed difference of 2.48 lies on the distribution, counting how many of the random permutations give differences higher or equal to 2.48, as well as lower or equal to –2.48, since the test is two-sided. There are 29 permutations outside these limits so the *p*-value is 29/10,000 = 0.0029, which is compatible with the *p*-value calculated initially for the regular *t*-test.

[1] See Appendix C for descriptions of the functions used, and the online R code.

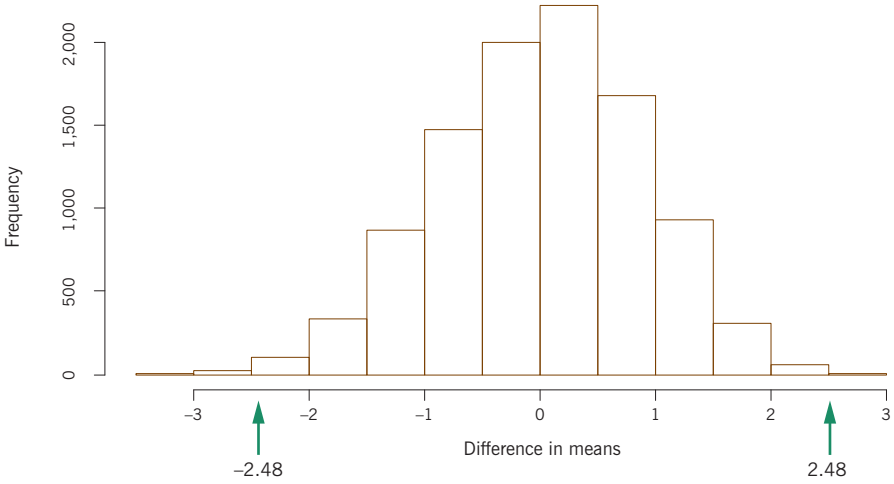


Exhibit 17.1:
Permutation distribution for test of difference in means of two populations based on samples of size 22 and 8. Of the 10,000 permutations 29 lie outside the limits of ± 2.48, hence the estimated p*-value is 0.0029*

To estimate the variability of the estimated difference correctly, without recourse to distributional assumptions, we would need repeated pollution samples of size 22 and 8 respectively from the populations of clay/sand and gravel locations from which the original data were obtained, which is clearly not possible since we only have one set of data. To simulate data from these two populations we can resort to *bootstrapping* the data. Samples are taken from the two sets of data, with replacement, which means that the same observation can be chosen more than once and some not at all. We do this repeatedly, also 10,000 times for example, each time computing the difference in means, leading to the *bootstrap distribution* of this difference, shown in Exhibit 17.2, alongside the separate bootstrap distributions of

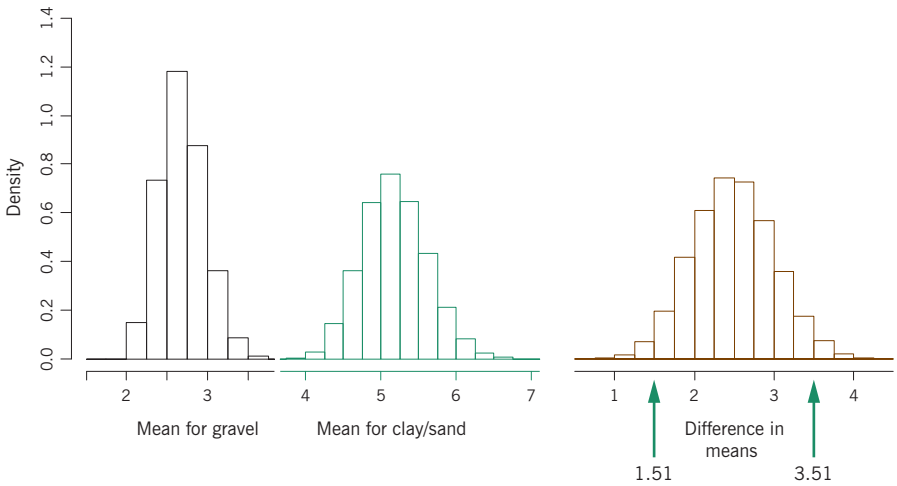


Exhibit 17.2:
Bootstrap distributions of the mean pollution for the gravel and clay/sand groups, based on 22 samples and 8 samples respectively, drawn with replacement 10,000 times from the original data. The right hand histogram is the bootstrap distribution of the differences, showing the limits for a 95% confidence interval

the two populations. The bootstrap distribution is an estimate of the true distribution of the differences, so to estimate the 95% confidence interval, we cut off 2.5% (i.e., 250 values out of 10,000) on each side of the distribution, obtaining the interval [1.51, 3.51]. This is more in line with the confidence interval obtained by the Welch method.

Test of association between two variables

Another common situation in statistical inference is to test an association, for example correlation, that is measured between two variables. In Chapter 1 we calculated a correlation between pollution and depth in the "bioenv" data set of –0.396 and a p-value of 0.0305 according to the two-tailed t-test for a correlation coefficient. This test relies on normality of the data but a distribution-free permutation test can also be conducted, as follows. Under the null hypothesis of zero correlation there is no reason to link any observation of depth with the corresponding observation of pollution in a particular sample, so we can randomly permute one of the data vectors. We do this 9,999 times, computing the correlation coefficient each time, and Exhibit 17.3 is the permutation distribution. The observed value of –0.396 is exceeded in absolute value by 315 of these randomly generated ones, and so the estimated p-value is 315/10,000 = 0.0315, almost the same as the t-test result.

Bootstrapping can be performed to obtain a confidence interval for the correlation coefficient. Now the pairs of depth and pollution values are kept together, and the sampling is done from their bivariate distribution by taking 30 samples at a time from the data set, with replacement (again, some samples are chosen more

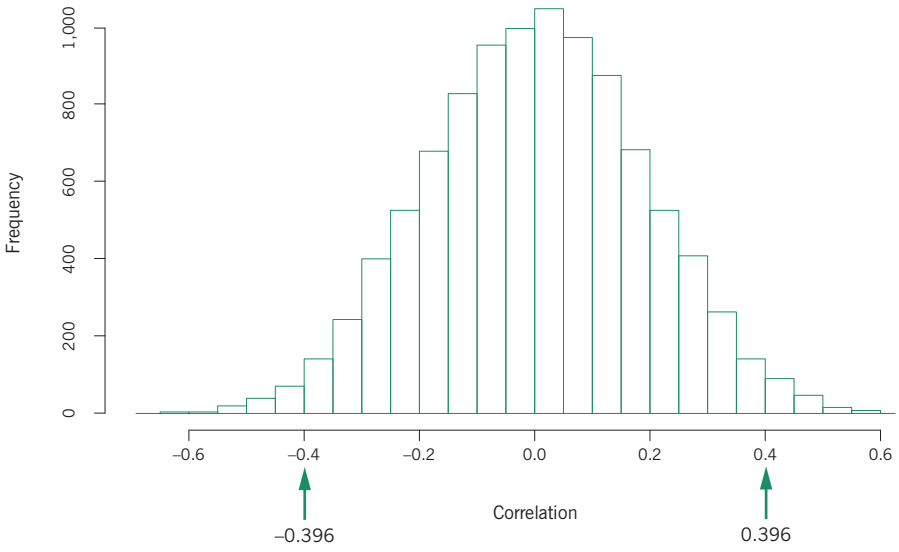


Exhibit 17.3: *Permutation distribution based on 9,999 estimates of the correlation between depth and pollution, under the null hypothesis of no correlation, together with the observed value of –0.396. The values ± 0.396 are indicated – there are 315 values equal to or more extreme, hence the* p*-value is 0.0315*

than once, some not at all). This is done 10,000 times, and each time a correlation coefficient is calculated, and Exhibit 17.4 shows their distribution. Cutting off 2.5% of the values on each tail gives a two-sided confidence interval for the correlation of [–0.704, –0.044]. Notice that the distribution in Exhibit 17.4 is not symmetric, and that this 95% confidence interval does not include zero, which is another way of saying that the observed correlation is significant at the 5% level of significance.

We have not exhausted all the possible alternative approaches in the last two sections. For example, a nonparametric Kruskal-Wallis rank test can be performed to test the difference in pollution between clay/sand and gravel samples, leading to a p-value of 0.0017. Or a Spearman rank coefficient can be computed between depth and pollution as –0.432 and its p-value is 0.021. Both these alternative approaches give results in the same ball-park as those obtained previously. Having shown these alternative ways of assessing statistical significance, based on statistical distribution theory with strong assumptions on the one hand, and using computationally intensive distribution-free methods on the other hand, the question is: which is preferable? It does help when the different approaches corroborate one another, but there is no correct method. However, we can eliminate methods that clearly do not fit the theory, for example normality-based methods should not be used when the data are clearly not normal. When we come to the multivariate case, however, the situation is much more complex, and in the absence of a theoretical basis for statistical testing, we rely more on the distribution-free approaches of permutation testing and bootstrapping.

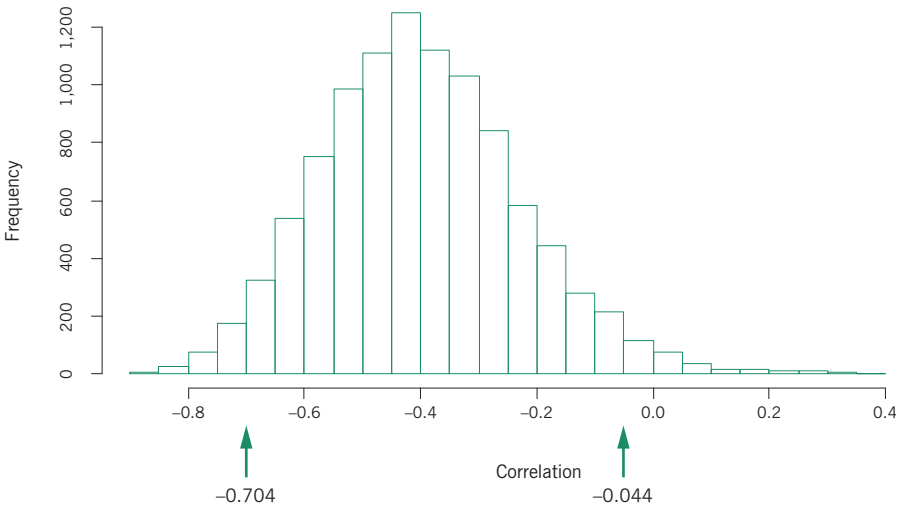
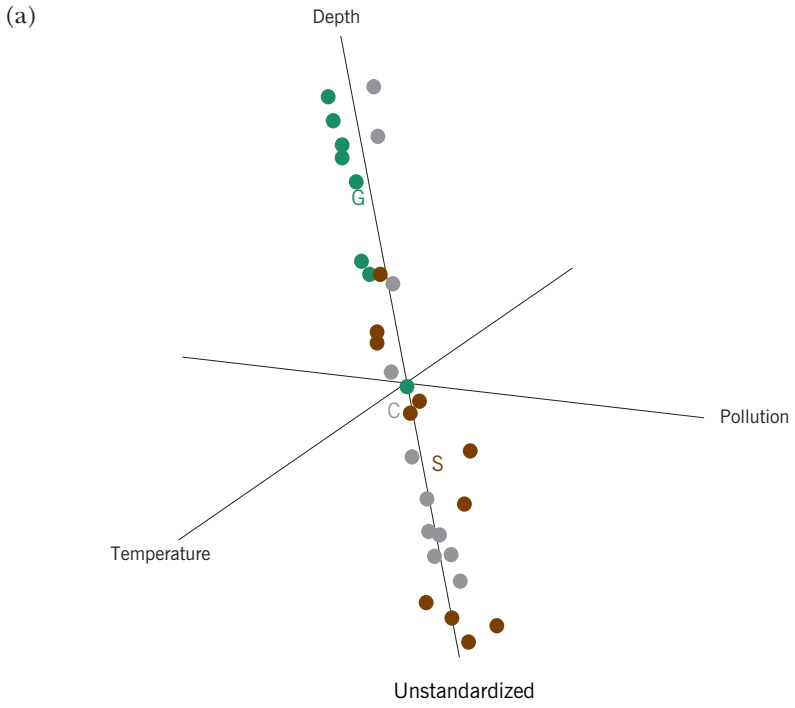


Exhibit 17.4:
Bootstrap distribution of the correlation coefficient, showing the values for which 2.5% of the distribution is in each tail

Exhibit 17.5: *Three-dimensional views of the 30 samples in the unstandardized (a) and standardized (b) Euclidean space of the three variables. Clay, sand and gravel samples are colour coded as gray, brown and green respectively, and their group average positions denoted by* C, S *and* G. *Since depth has a much higher range of numerical values than the other two variables, it would dominate the computation of inter-group difference if the data were not standardized in some way*



Multivariate test of group difference

Suppose now that we wanted to test the differences between the sediment groups based on all three continuous environmental variables depth, pollution and temperature. This time let us also keep the three sediment groups separate. Even in the univariate case, when we pass to more than two groups, the notion of a negative difference no longer exists – any measure of difference will be a strictly positive number. Furthermore, when we pass to more than one variable then the issue of standardization is crucial in the measure of group difference, since the variables have to contribute equitably to the measure of difference. To rub this point in even further, consider the positions of the samples in unstandardized and standardized coordinates in Exhibit 17.5. The centroids of the three sediment groups are also shown, and it is clear that standardization is necessary, otherwise depth would dominate any measure of intergroup difference. We are going to measure the difference between the three sediment groups by the lengths of the sides of the triangle in the standardized space – see the triangle in Exhibit 17.5(b). If these lengths are large then the group means are far apart, if they are small then the means are close together. The question then is whether they are significantly far apart. The sum of these three lengths turns out to be 6.303. To obtain a p-value a permutation test is performed by randomly allocating the C, S and G labels to the data samples many times, and each time computing the same statistic, the sum of the distances between group means. The permutation distribution is shown in Exhibit 17.6, and the observed statistic lies well into the tail of the distribution, with a p-value of 0.0032. Notice that now it is only the right tail that is counted, since the value of 0 on the left side of the distribution indicates the null hypothesis of no difference.

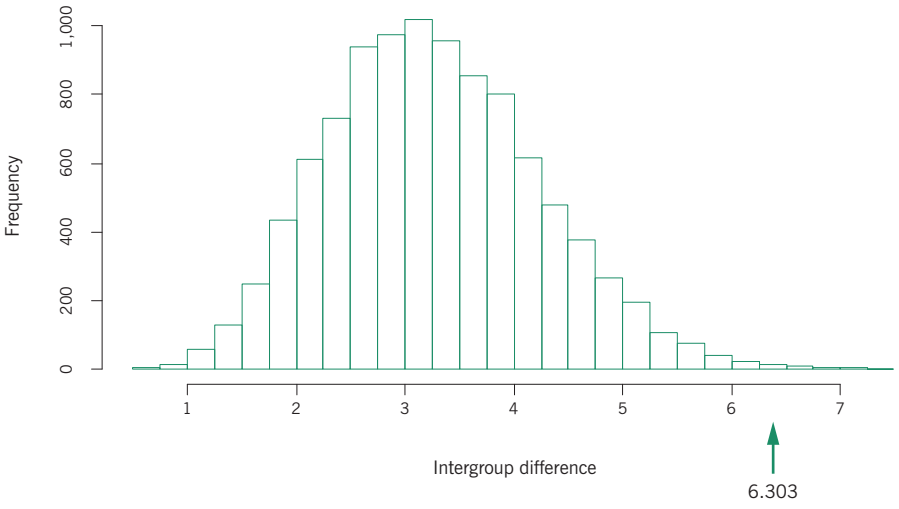
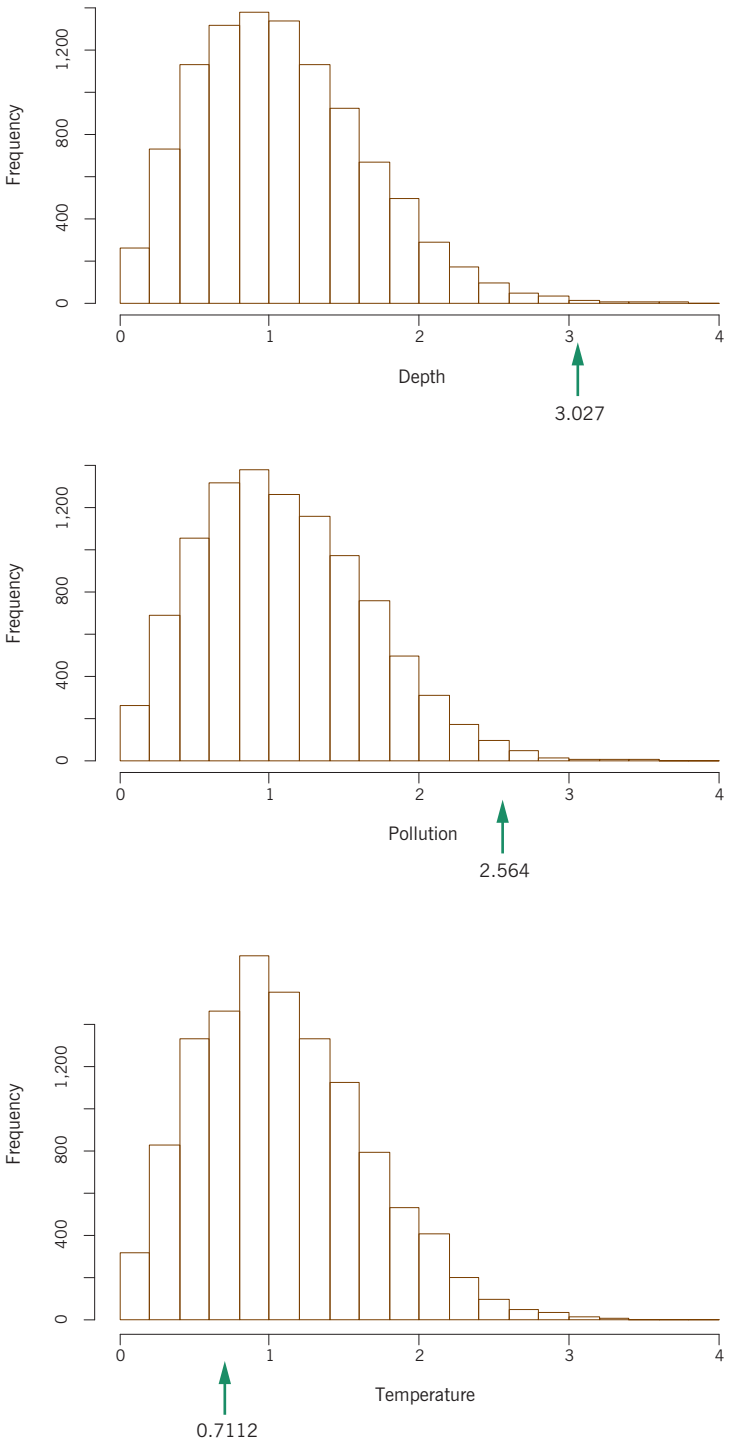


Exhibit 17.6:
Permutation distribution of measure of intergroup difference in standardized multivariate space. There are 32 of the simulated values greater than or equal to the observed value of 6.303, hence the p*-value is 32/10,000 = 0.0032*

Exhibit 17.7: *Permutation distributions for measure of intergroup difference based on single variables. The observed difference is indicated each time and the* p*-values are 0.0032, 0.0084 and 0.7198 respectively.*



Having concluded that the group differences are significant there are two further aspects to be considered, since there are three variables and three groups involved: first, are all groups significantly different from one another? and second, which variables contribute mostly to the difference? These two questions are related, since it may be that a group may be different from another on only one or two of the variables, whereas the third group may be different from the other two on all three variables. Anyway, let us consider the latter question first: are the groups significantly different on all three variables? We can perform the same test three times using one variable at a time – here it would not matter if the data were standardized or not, but we continue to use the standardized form since it is easier to compare the three results. Exhibit 17.7 shows the three permutation distributions and it is clear that temperature is not at all different between the three sediment groups, so we could drop it from further consideration.

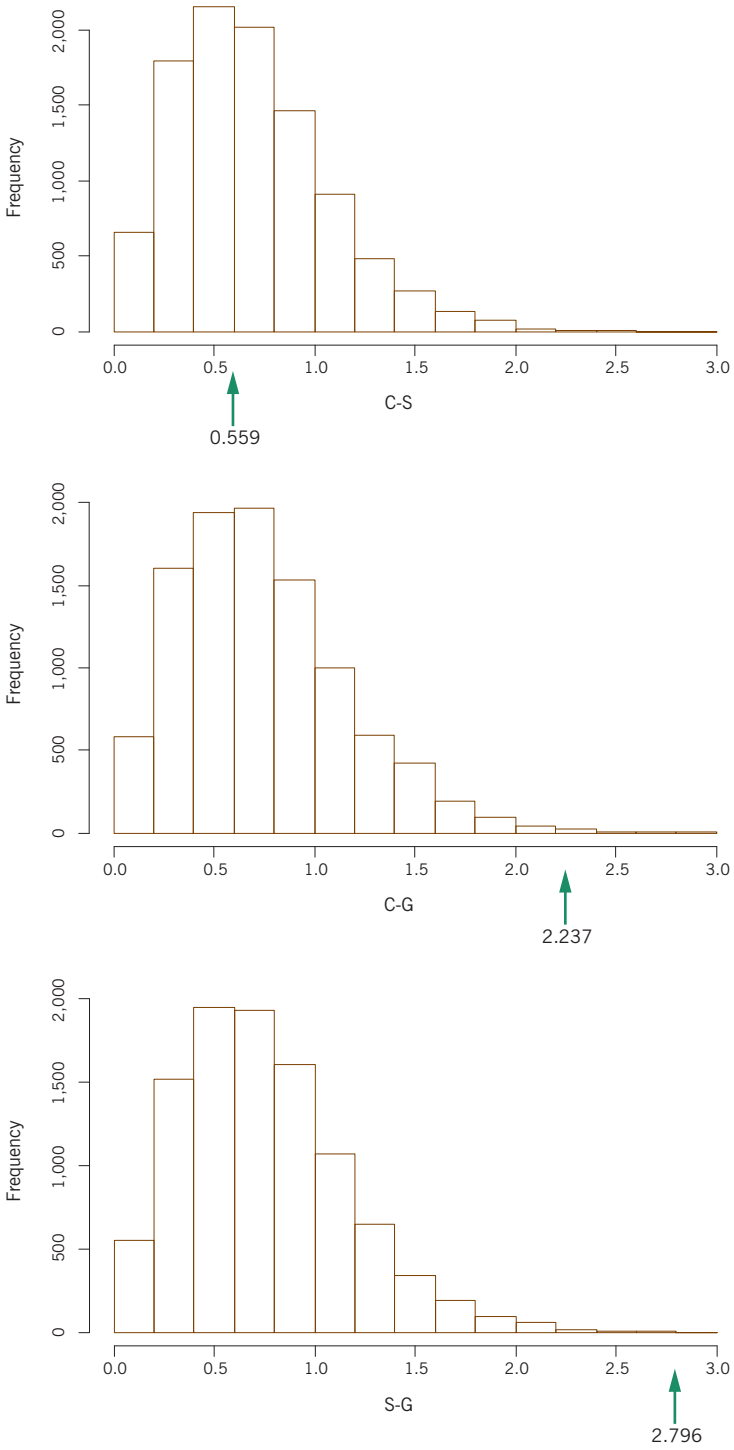
Next, we examine whether the groups are all different from one another, based on just depth and pollution. The differences between C and S, S and G and C and G are computed, similar to the multiple comparison procedure in ANOVA, and their permutation distributions generated, shown in Exhibit 17.8. It is clear that there is no significant difference between clay and sand groups (hence our aggregating them in the initial example of this chapter), whereas they are both highly significantly different from the gravel group.

Test of association between groups of variables

The multivariate equivalent of testing a correlation coefficient is when there are several predictor variables being used to explain one or more response variables. The most relevant case to us is in canonical correspondence analysis (CCA), when many biological variables are being related to several environmental variables, for example, via a process of dimension reduction in the biological space. There are two ways to assess this relationship: one way is to simply include all the environmental variables in the model and test for their overall significance, while another more laborious way is to look for a subset of significant environmental predictors, eliminating the insignificant ones. We shall illustrate these two strategies again using the simple "bioenv" data set, leaving a more substantial application to the case study of Chapter 19.

The inertia of the biological data in this example (species *a*, *b*, *c*, *d* and *e*) is 0.544 (see Chapter 13). When using depth, pollution and temperature as environmental predictors in a CCA, the inertia accounted for is 0.240, or 44.1%. We can generate a permutation distribution to test whether this percentage is significant. As in the case of the correlation of two variables, under the null hypothesis of no relationship, the biological and environmental data vectors can be randomly paired, keeping the biological vectors (with

Exhibit 17.8:
Permutation distributions for measure of pairwise intergroup differences based on depth and pollution. The observed difference is indicated each time and the p*-values are 0.5845, 0.0029 and 0.0001 respectively*



five abundance values) and the environmental vectors (with three continuous measurements) intact. If we do this 9,999 times, we do not get any case that improves the figure of 44.1% inertia explained, so the p-value is 0.0001, highly significant.

Dropping one variable at a time and repeating this exercise, we obtain the following percentages of explained inertia and p-values for the three different pairs of variables:

Depth and pollution:	42.7% ($p = 0.0001$)
Depth and temperature:	11.4% ($p = 0.1366$)
Pollution and temperature:	37.4% ($p = 0.0001$)

and for a single variable at a time:

Depth:	10.0% ($p = 0.0366$)
Pollution:	36.3% ($p = 0.0001$)
Temperature:	1.1% ($p = 0.8642$)

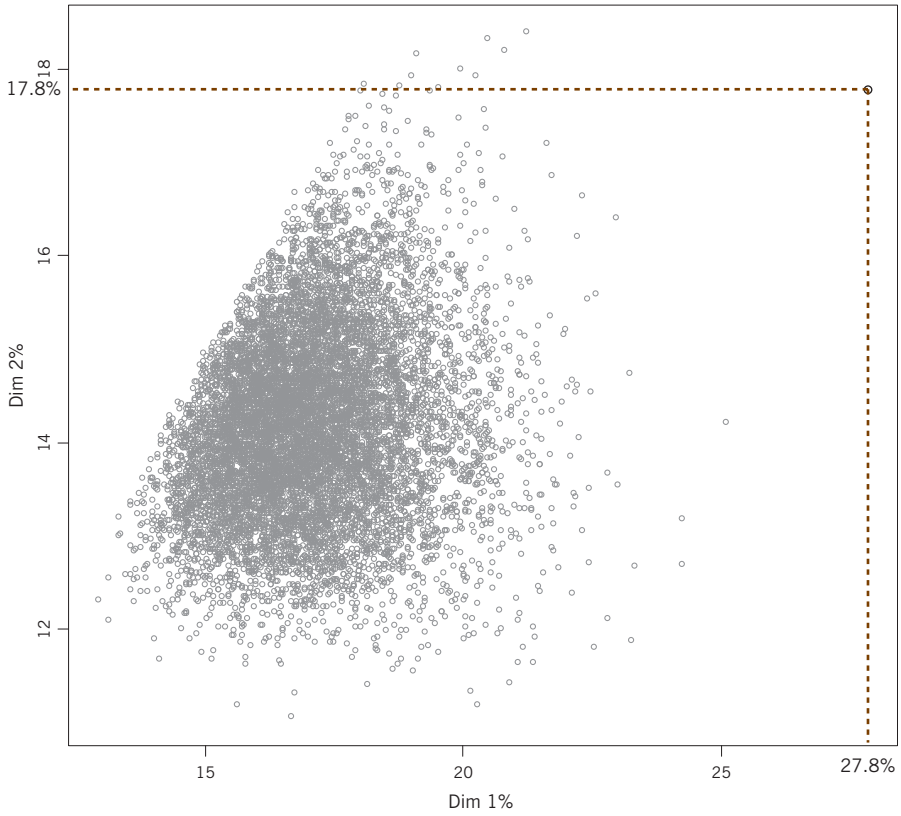
So it seems clear that temperature has no predictive role and can be dropped. Pollution is the best predictor and if a forward stepwise process were followed, then pollution would be the first to enter the model. The only question remaining is whether depth adds significantly to the model that is driven mainly by pollution. This can be tested by generating a permutation distribution with pollution as a predictor, unpermuted, while just the depth values are permuted randomly. After the usual 9,999 permutations of the depth vector, the result is that the percentage of inertia explained by depth and pollution, seen above to be 42.7%, is the 222nd highest value in the sorted list of 10,000, so the p-value for the additional explained inertia of depth is 0.0222, significant at the 5% level. The final model would thus include pollution and depth.

Other permutation tests for ordination

In any dimension-reduction technique to establish an ordination of a data set, the objective is to separate what is "signal", that is true structure, from "noise", that is random variation. In Chapter 12 we discussed an informal way of judging which dimensions are "significant" from the appearance of the scree plot (see Exhibit 12.6 and related description). A permutation test can make a more formal decision about the dimensionality of the solution. In a PCA, the correlations between the variables combine to form the principal

axes of an ordination, so if there is no correlation between the variables then there is no structure. Hence a permutation test for the principal axes of the "climate" data set can be performed by generating several random data sets under the null hypothesis of no correlation between the variables by randomly permuting the values down the columns of each variable. The eigenvalues of these randomly generated data sets yield permutation distributions of the eigenvalues under the null hypothesis. Since the total variance in a PCA of standardized data is a fixed number, it is equivalent to look at the percentages of variance explained on the axes. Exhibit 17.9 shows the scatterplot of the first and second percentages of variance for the 9,999 permuted data sets, along with the actual values of 27.8% and 17.8% in the original data set. The p-values are again calculated by counting how many of the values are greater than or equal to the observed ones, only 1 for the first dimension (the observed value itself) and 13 for the second, hence the p-values are 0.0001 and 0.0013 respectively. Continuing with the third and higher dimensions, the p-values are 0.0788, 0.2899, 0.9711 and so on, none of which is significant.

Exhibit 17.9:
Scatterplot of percentages of variance on the first two dimensions of 10,000 PCAs, one of which is based on the observed data set "climate" and the other 9,999 are computed using random matrices obtained by permutation



Hence, the two-dimensional solution, accounting for 45.6% of the variance, is considered the appropriate solution, and the remaining 54.4% of the variance is regarded as random variation.

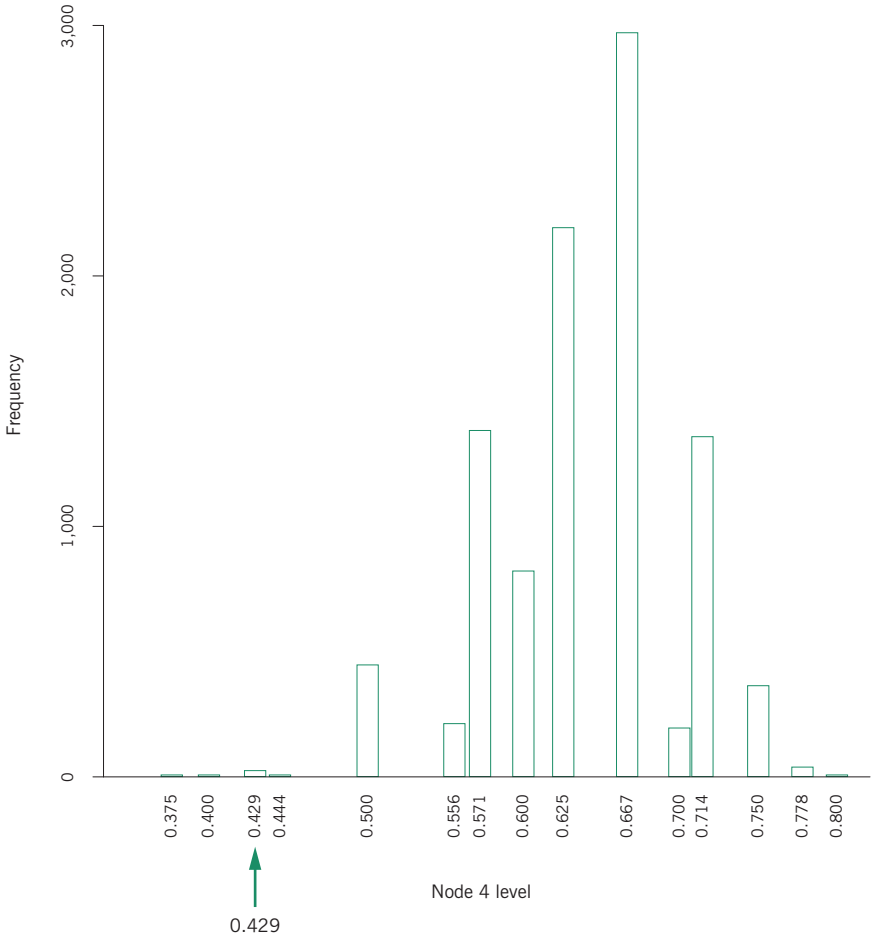
The situation is slightly different for methods like CA, LRA and CCA that do not have a fixed total variance, and have weights attached to the rows and columns. Whereas in PCA it is equivalent to consider the eigenvalues or their percentages relative to the total variance, in these other methods such as CA, for example, the total inertia of an abundance matrix can be very high compared to the variance of the permuted matrices under a null model of no relationship between the species. So we would base our decision about significance of the dimensions purely on the percentages of inertia. An example will be given in the case study of Chapter 19.

A permutation test for clustering

The same issues arise when performing a cluster analysis: for example, which are the "significant" clusters, or in a hierarchical clustering at what level should the dendrogram be cut for the identification of "significant" clustering? This is a difficult question and we present one possible approach to identifying a significant cutpoint. The levels of the nodes at which clusters are formed are first saved for the original dendrogram, for example in the dendrogram on the right of Exhibit 7.8, based on clustering zero/one data using the Jaccard index, the levels of the nodes are (from the bottom up): 0.200, 0.250, 0.333, 0.429, 0.778 and 1.000. Now we generate a permutation distribution for these levels by randomly permuting the columns of the data matrix, given in Exhibit 5.6, so that we have a large number of simulated values (again, 9,999) under the hypothesis of no relationship between the species. For each permuted matrix the node levels are stored, and then for each level we count how many are less than or equal to the originally observed node level. For significant clustering we would expect the node level to be low. The p-values associated with each node are (again, from bottom up): 0.1894, 0.0224, 0.0091, 0.0026, 0.7329, 1.000, so that node level 4, which cuts the sample into three clusters, is the most significant. Exhibit 17.10 shows the permutation distribution for the node 4 levels and the observed value of 0.429. There are only 26 permutations where the node level is lower than or equal to 0.429, hence the p-value of 0.0026.

As a contrasting example, the same strategy was applied to the dendrogram of Exhibit 7.10 that clusters 30 samples based on their standardized Euclidean distances using variables depth, pollution and temperature. None of the p-values for the 29 nodes in this example are less than 0.05, which indicates that there are no real clusters in these data, but rather a continuum of dispersion in multivariate space.

Exhibit 17.10: *Permutation distribution of node 4 levels, corresponding to a three-cluster solution, for the presence-absence data of Exhibit 5.6 – see the dendrogram on the right of Exhibit 7.8. There are 26 permutations (out of 10,000) that are less than or equal to 0.429, the value of the level in the dendrogram*



SUMMARY: Inference in multivariate analysis

1. Conventional statistical inference relies on assuming an underlying distribution of the data under a null hypothesis (e.g., a hypothesis of no difference, or of no correlation), called the *null distribution*. The unusualness of the observed value (e.g., a difference or a correlation) is then judged against the null distribution and if its probability of occurring (i.e., *p*-value) is low, then the result is declared statistically significant.

2. Distribution-free methods exist that free the analyst from assuming a theoretical distribution: null distributions can be generated by permuting the data under the null hypothesis, and the variability of observed statistics can be estimated using bootstrapping of the observed data.

3. In the multivariate context, where theory is much more complex, we shall generally rely purely on computer-based permutation testing and bootstrapping.

4. To assess the significance of group differences, the null hypothesis of no difference implies that we can allocate observations to any groups. A statistic measuring group difference is postulated, and then the same statistic is measured on data that have been randomly permuted a large number of times by randomly assigning the group affiliations to the observations. The statistic computed on the original data is then judged against the permutation distribution to obtain a *p*-value.

5. To assess the association between two sets of variables, a statistic measuring this association is first postulated. Under the null hypothesis of no difference we can randomly connect the first and second sets of data, and doing this many times generates a null distribution of the association measure. The observed association measured on the original data is once again judged against the permutation distribution to obtain a *p*-value.

6. The parts of variance/inertia, or eigenvalues, can also be assessed for statistical significance by generating a null distribution of their percentages of the total, under an hypothesis of no relationship between the variables (usually columns of the data matrix), in which case the values for each variable can be permuted randomly to generate a null distribution of each eigenvalue.

7. We propose a similar procedure for hierarchical cluster analysis, where clusteredness is indicated by low node levels. The data for each variable are permuted randomly and each time the same clustering algorithm performed, generating a permutation distribution for each level under the null hypothesis. Observed node levels that are in the lower tail of these permutation distributions will indicate significant clustering.

List of Exhibits