MDS can be based on any of the measures of dissimilarity described in Chapter 15 but is not restricted to these. For example, Guiller *et al.* (1998) calculated genetic dissimilarities (Nei's and Rogers' indices) between 30 North African populations of the snail *Helix aspersa*, based on 17 enzyme loci. They used MDS to examine the relationships between the populations.

We will illustrate MDS using some recent data sets from the biological literature.

**Genetic structure of a rare plant**

In Chapter 15, we described the work of McCue *et al.* (1996), who measured the genetic structure of a rare annual plant (*Clarkia springvillensis*) in California. They identified eight subpopulations and calculated Cavalli-Sforza genetic distances between subpopulations from isozyme analysis of tissue samples. We will use their genetic distances as dissimilarities and examine the relationships between the subpopulations using MDS.

**Habitat fragmentation and rodents**

In Chapter 13, we introduced the study of Bolger *et al.* (1997) who surveyed the abundance of seven native and two exotic species of rodents in 25 urban habitat fragments and three mainland control sites in coastal southern California. Besides the variables representing the species, other variables recorded for each fragment and mainland site included area (ha), percentage shrub cover, age (years), distance to nearest large source canyon and distance to nearest fragment of equal or greater size (m). We will first calculate dissimilarities in species composition between the 25 fragments and three mainland sites and use MDS to represent the relationship between these objects. We will then examine relationships with other fragment characteristics such as area for the 25 fragments.

**Geographic variation and forest bird assemblages**

Mac Nally's (1989) study on forest birds was first used in Chapter 17. The data set consisted of the maximum abundance (from four seasons) for 102 species of birds for 37 sites in southeastern Australia. These sites were actually replicates of five different forest types, four each of Gippsland manna gum, montane forest, foothills woodland, box-ironbark and river redgum with the remaining 17 sites not able to be classified into one of the habitats. An obvious question is whether the five habitat types are different in the composition of their bird assemblages.

### 18.1.1 Classical scaling – principal coordinates analysis (PCoA)

Principal coordinates analysis (PCoA) is closely related to PCA (Chapter 17) and is sometimes called classical scaling. We will only provide a brief introduction to PCoA here (see Legendre & Legendre 1998 for complete details), mainly because it is not used that much as a scaling (ordination) technique in biology. The steps in PCoA are as follows.

- Create an $n$ by $n$ matrix of dissimilarities between objects ($d_{hi}$), based on any of the dissimilarity measures described in Chapter 15.
- Transform these dissimilarities to $-0.5d_{hi}^2$. This transformation maintains the original dissimilarities during subsequent calculations (Legendre & Legendre 1998).
- These transformed dissimilarities are double centered by subtracting the means for the relevant row and column and adding the overall mean from the dissimilarity matrix. This centering removes the first, and trivial, eigenvector in the next step. The relative positions of the objects in the final configuration won't be affected by the double centering.
- This symmetric $n$ by $n$ matrix of transformed dissimilarities is then subjected to a spectral decomposition to obtain the eigenvectors and their eigenvalues, in exactly the same way as we treated a matrix of associations (covariances or correlations) between variables in a *R*-mode PCA. Most of the information (as measured by the eigenvalues) in the dissimilarity matrix will be in the first few eigenvectors (Box 18.1).
- As with PCA (Chapter 17), the eigenvectors are scaled, usually by the square roots of the eigenvalues (Legendre & Legendre 1998).
- The coefficients of these eigenvectors are then used to position the objects relative to each other on the scaling plot (Figure 18.1).



## Box 18.1 Worked example of PCoA: habitat fragmentation and rodents

We will use the data on rodent numbers from 25 canyon fragment and three mainland sites in California from Bolger *et al.* (1997) to illustrate PCoA. Because the sites were very different in size, we standardized the total abundance for each site to range between zero and one and calculated a matrix of Bray–Curtis dissimilarities between the sites. This matrix was then used for the PCoA. Of the 28 possible eigenvectors, ten had zero eigenvalues and seven had negative eigenvalues but nearly 90% of the variance was explained by the first two components so only these were used for the scaling plot of sites.

| | Axis 1 | Axis 2 |
|---|---|---|
| Eigenvalues | 5.255 | 1.724 |
| Percentage variation | 66.081 | 21.681 |
| Cumulative percentage variation | 66.081 | 87.762 |

The PCoA scaling plot of the 28 sites based on the original Bray–Curtis dissimilarities of data range standardized by site is shown in Figure 18.1. When corrected for total abundance at a site, the three mainland sites were almost identical and were not distinguishable from most of the canyon fragments. Acuna, El Mac and 54th Street separated from the other sites, especially along axis 2. These three sites also stood out from the others in the scaling plot from a CA of these 28 sites (Chapter 17, Figure 17.5). The agreement with CA is because the latter emphasizes proportional abundance of species at each site, as does the PCoA when the dissimilarity is calculated on abundances standardized to the same maximum value at a site. Note, however, that the CA did separate the three mainland sites from each other, a pattern not observed in the PCoA, probably reflecting differences in the sensitivity of the two dissimilarity measures (chi-square and Bray–Curtis) to changes in proportional abundance.

If the original data were centered by variable means and Euclidean distance was used to create the matrix of dissimilarities between objects, the relative positions of objects in the PCoA scaling will be similar to those for the scaling plot from a PCA based on a matrix of covariances between variables. If the original data were double transformed by row and column totals so that chi-square distance was used to create the dissimilarity matrix, the relative positions of objects in the PCoA scaling will be similar to those for the scaling plot from a CA. So PCoA can be viewed as a generalization of PCA that allows a much wider range of dissimilarity measures to be used.

Another way of viewing PCoA is a translation of dissimilarities between objects into Euclidean distances, the actual distances between objects in multidimensional space (Legendre & Anderson 1999a). If the original dissimilarities were metric (such as Euclidean or chi-square), and all eigenvectors are retained, then the distances in principal coordinate space are the same as the original dissimilarities because all the variance in the original dissimilarity matrix is retained in the principal coordinates. In contrast, biologists often use non-metric dissimilarities, like Bray–Curtis for species abundance data, and the principal coordinates represent only part of the variation in the original dissimilarities. Unfortunately, the remainder may be represented by negative eigenvalues, which are very difficult to interpret. This may not be a problem if we are using PCoA as a variable reduction technique because the first few eigenvalues will be positive. However, if we wish to use all the principal coordinates derived from a non-metric dissimilarity matrix, such as in

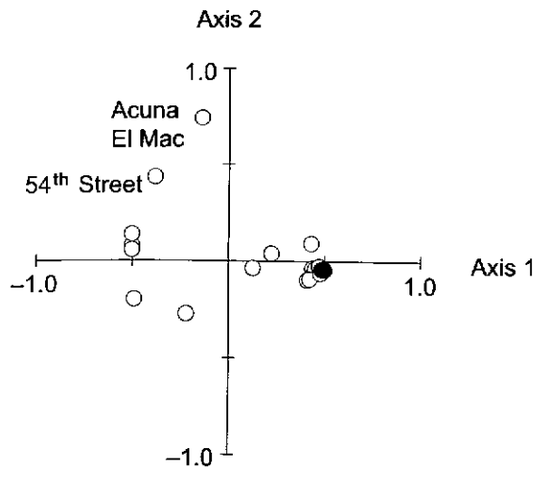



**Figure 18.1** PCoA scaling/ordination plot of the 28 sites from Bolger *et al.* (1997) based on a Bray–Curtis matrix of dissimilarities between sites, standardized so all sites have maximum abundance of one. The three mainland sites are filled symbols.

distance-based redundancy analysis (db-RDA; see Section 18.1.3), then we usually have to correct for the negative eigenvalues. These corrections are somewhat technical (Legendre & Legendre 1998, Legendre & Anderson 1999a) and may result in conservative tests of complex hypotheses (McArdle & Anderson (2001).

When dealing with species abundance data, Minchin (1987) showed that the scaling plots of sampling units produced by PCoA could distort underlying ecological gradients. In particular, PCoA would force long gradients (i.e. with considerable species turnover from one end to the other) into curved patterns in the configuration in second and higher dimensions. This distortion occurred even when more robust dissimilarity measures like Bray–Curtis were used and Minchin (1987) argued that this was because PCoA, like PCA, is based on a linear relationship between dissimilarity and ecological distance, whereas the relationship was nonlinear, particularly for large dissimilarities. Also, PCoA does not provide a simple way of interpreting the new coordinates in terms of the original variables (Legendre & Legendre 1998). Although these problems do not rule out PCoA as a scaling technique for other types of data, biologists don't use PCoA very much by itself because modern desktop computers make enhanced scaling techniques (Section 18.1.2) so accessible. However, PCoA was used by Rundle & Jackson (1996) who measured the abundance of 15 species of littoral zone fish from five sites in each of three lakes in Ontario, Canada. They calculated Bray–Curtis dissimilarities between the 15 sites and then subjected the dissimilarity matrix to a PCoA. The first two axes explained over 69% of the variation in the original dissimilarity matrix and one lake clearly separated from the other two along the first axis.

We illustrate the use of PCoA on the data from Bolger *et al.* (1997), who recorded the abundance of nine species of rodents in 25 habitat fragments and three mainland sites in southern California – see Box 18.1. We calculated a matrix of Bray–Curtis dissimilarities between sites. Close to 90% of the variation was explained by the first two axes.

## 18.1.2 Enhanced multidimensional scaling

### Enhanced algorithm

Methods for MDS more familiar to biologists involve additional steps, beyond the initial scaling used by PCoA, to improve the fit between the observed dissimilarities between objects ($d_{hi}$) and the inter-object distances in the configuration ($\tilde{d}_{hi}$). Jackson (1991) termed these methods "enhanced multidimensional scaling". Basically, these methods iteratively reposition the objects in the configuration using an algorithm that improves the fit between the dissimilarities and the inter-object distances, the latter measured by a form of Minkowski metric such as Euclidean distance. The most commonly used algorithm for enhanced MDS is KYST, developed from methods first proposed by Kruskal (1964a,b), although some software offers the alternative ALSCAL program. The approach is surprisingly simple, although the computations would be very tedious without computer software. The steps for an enhanced MDS are as follows (Figure 18.2).

1. Set up a data matrix and make decisions about transformations or standardizations of the data.

2. Calculate a matrix of dissimilarities between objects ($d_{hi}$) using any of the dissimilarities described in Chapter 15. Similarities could also be used; it makes no difference in the subsequent steps.



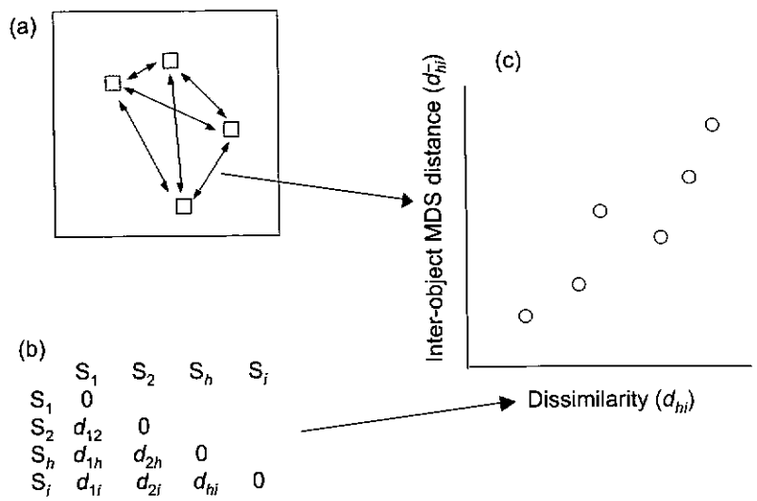



**Figure 18.2** Illustration of the links between (a) configuration (scaling/ordination) plot, (b) dissimilarity matrix and (c) Shepard plot in enhanced MDS. $S_1$, $S_2$, etc., are objects, e.g. sampling units.

3. Decide on the number ($k$) of dimensions (i.e. axes) for the scaling, which will be a compromise between the need to get the fit between dissimilarities and inter-object distances as good as possible and minimizing the number of scaling dimensions for simple interpretation.

4. Arrange the objects in a starting configuration in the $k$-dimensional space (i.e. on the plot), either at random or more commonly using coordinates from a PCoA or even a PCA.

5. Move the location of objects in the $k$-dimensional space iteratively so that at each step, the match between the inter-object distances in the configuration ($\tilde{d}_{hi}$) and the actual dissimilarities ($d_{hi}$) improves. The iterative procedure uses the method of steepest descent (see Kruskal 1964a,b for details).

6. The final position of the objects and therefore the final configuration plot is achieved when further iterative moving of the objects can no longer improve the match between the inter-object distances in the configuration and the actual dissimilarities.

We can show the relationship between inter-object distance and dissimilarity for all pairs of objects in a Shepard diagram, which is simply a scatterplot with dissimilarity ($d_{hi}$) on the horizontal axis and inter-object distance ($\tilde{d}_{hi}$) on the vertical axis (Figure 18.2(c)). Now consider a linear or nonlinear regression model relating inter-object Euclidean distance ($\tilde{d}_{hi}$) as the response variable to dissimilarity ($d_{hi}$) as the predictor variable. The differences between the observed inter-object distances and those predicted by the regression model ($\hat{\tilde{d}}_{hi}$, sometimes termed "disparities" in the MDS literature) are the residuals from the regression model. These residuals can be used to measure the match between the calculated dissimilarities and the inter-object distances in the configuration.

One measure of fit is Kruskal's stress:

$$\sqrt{\frac{\sum (\tilde{d}_{hi} - \hat{\tilde{d}}_{hi})^2}{\sum \tilde{d}_{hi}^2}} \qquad (18.1)$$

In Equation 18.1, the summation is over all possible $n(n-1)/2$ pairwise distances and dissimilarities. If there is a perfect metric match between inter-object distance and dissimilarity (i.e. they are directly proportional to each other), then the residuals and stress will be zero. The lower the stress value, the better the match. There are other versions of stress used to measure fit (e.g. see Jackson 1991) and it is important to know which your software uses because they are scaled, and therefore interpreted, differently. The version in Equation 18.1 is the one usually incorporated in the KYST algorithm and most commonly used by biologists. When stress is based on a parametric linear or nonlinear regression model relating inter-object distances to dissimilarities, we have metric MDS.

It is common for the Shepard plot to show a nonlinear relationship between inter-object distance and dissimilarity (Figure 18.3(b)). While this

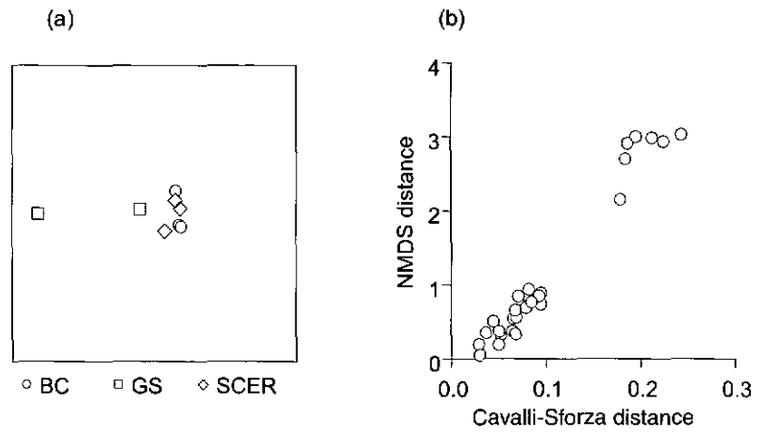



**Figure 18.3** (a) NMDS scaling/ordination plot of the eight subpopulations of the plant Clarkia springvillensis based on Cavalli–Sforza genetic distances between subpopulations; from McCue et al. (1996). (b) Shepard plot showing the relationship between Cavalli–Sforza genetic distances between subpopulations and NMDS distances between subpopulations.

might suggest that a nonlinear model relating inter-object distance and dissimilarity is most appropriate, a more robust approach is to fit a monotonic regression. This is a form of nonparametric regression that relates the rank orders of the two variables (Chapter 5). So stress now measures the concordance in the rank order of the observed inter-object distances and those predicted from the dissimilarities. When stress is based on rank orders, we have non-metric MDS (NMDS).

A third type of MDS has been developed by Faith *et al.* (1987) and is termed hybrid MDS (HMDS). They noted that for species abundance data, sampling units at the ends of long ecological gradients often have few or no species in common and this can result in the nonlinear relationship between dissimilarity and inter-object ("ecological") distance mentioned in the previous paragraph. Importantly, it seemed that a linear relationship between dissimilarity and inter-object distance was appropriate for small dissimilarities but inappropriate for larger dissimilarities. Their hybrid approach generates two dissimilarity matrices. The first deletes dissimilarities above a threshold value and then uses a metric (linear) MDS to measure stress. The second matrix uses all the dissimilarities and uses a non-metric MDS to measure stress. The final configuration is the one that minimizes the combination of the two stress values. The choice of dissimilarity threshold is a difficult one, with Faith *et al.* (1987) originally proposing 0.8 (for Bray–Curtis or Kulczynski dissimilarities) but also suggesting that some continuous function could also be used. Our experience is that HMDS does not offer much advantage over NMDS, even for ecological data sets, and is only available in specialized software anyway.

### Interpretation of final configuration

We illustrate the use of NMDS with the data set on genetic differences between subpopulations of a species of plant from McCue *et al.* (1996) in Box 18.2, the habitat fragmentation study of Bolger *et al.* (1997) in Box 18.3 and the forest bird community study from Mac Nally (1995) in Box 18.4. The final configuration is the scatterplot of objects in a scaling or ordination diagram (Figure 18.3, Figure 18.4, Figure 18.5). The interpretation of this plot depends on how good a representation it is of the actual dissimilarities, i.e. how low the stress value is. Clarke (1993) provided some guidelines for stress values based on ecological (species abundance) data. Stress values greater than 0.3 indicate the configuration is no better than arbitrary and we should not try and interpret configurations unless stress values are less than 0.2, and ideally less than 0.1. These thresholds are for Kruskal's stress formula in Equation 18.1, while some software may use different versions that require different guidelines. We can always reduce the stress value, i.e. improve the fit between dissimilarities and inter-object distances, by increasing the number of dimensions in the scaling. However, the more dimensions we use, the more difficult the display and interpretation of the final configuration, so we are trying to achieve a compromise



## Box 18.2 Worked example of enhanced MDS: genetic structure of a rare plant

McCue *et al.* (1996) sampled eight subpopulations of the rare annual plant (*Clarkia springvillensis*) from three sites along the Tule River in California. Two sites, Bear Creek (BC) with three subpopulations and the Springville Clarkia Ecological Reserve (SCER) with three subpopulations, were separated by about 300 m and the third site, Gauging Station (GS) with two subpopulations, is approximately 8 km apart. The non-metric MDS algorithm produced identical configurations from all random starts and the stress of the final configuration was 0.045, indicating that the scaling/ordination of the subpopulations closely matched the Cavalli–Sforza genetic distances between the subpopulations. The final scaling plot of the subpopulations (Figure 18.3) indicates that the two Gauging Station subpopulations are genetically different from the remaining subpopulations, with subpopulation GS1 being the most distinct.

## Box 18.3 Worked example of enhanced MDS: habitat fragmentation and rodents

We will use the data on rodent numbers from 25 canyon fragment and three mainland sites in California from Bolger *et al.* (1997) to illustrate NMDS. Because the sites were very different in size, the data were standardized so that each site had a maximum total abundance of rodents of one. We were interested in comparing sites based on species composition and abundance but without patterns being confounded by very different areas.

A matrix of Bray–Curtis dissimilarities between all 28 sites was calculated and subjected to non-metric MDS. From 20 random starts in two dimensions, the minimum stress value of 0.054 was achieved from four starts, although all 20 starts produced very similar final configurations, one of which is displayed in Figure 18.4, with a small range of stress values (0.054–0.059). The mainland sites were not clearly separate from the fragments and the pattern of sites was similar to that in the PCoA plot. The same fragment sites were close to the mainland sites and Acuna, El Mac and 54th Street were most different to the mainland sites (Figure 18.4). It is interesting to compare the pattern from the NMDS to that from the CA on the same data described in Chapter 17 (Figure 17.5). Although the distances between the sites are different in the two plots, the broad pattern of Acuna, El Mac and 54th Street being separate was consistent in both analyses.

Correlations were calculated between the two dimensional configuration (scores) of sites and each of the six habitat variables (total area, shrub area, percentage area of shrubs, distance to nearest large source canyon and distance to nearest fragment of equal or greater size, age). Randomization testing showed that only percentage of shrub was significantly related to the configuration of sites, although the result for age suggested a pattern worth investigating further.

| Variable | *n* | *r* | *P* |
|---|---|---|---|
| Area | 28 | 0.28 | 0.380 |
| Shrub | 28 | 0.33 | 0.250 |
| Percentage shrub | 28 | 0.69 | 0.010 |
| Distance nearest source | 25 | 0.18 | 0.740 |
| Distance nearest fragment | 25 | 0.20 | 0.640 |
| Age | 25 | 0.47 | 0.050 |

## Box 18.4 Worked example of enhanced MDS: geographic variation and forest bird assemblages

The data set from Mac Nally (1989) consisted of the maximum abundance (from four seasons) for 102 species of birds for 37 sites in southeastern Australia. A matrix of Bray–Curtis dissimilarities between sites was constructed. No standardization was used because the data were densities of birds, rather than absolute counts. This means that species with high densities will dominate the dissimilarities between sites. A non-metric MDS in two dimensions, using 20 random starts, resulted in a stress value of 0.14. Using three dimensions, a stress value of 0.08 was achieved from 12 of the 20 random starts, so the three dimensional solution was used. The scaling/ordination plot of the 37 sites in the first two of the three dimensions (Figure 18.5(a)) showed clear separation of sites dominated by Gippsland Manna Gum and River Red Gum, and to a lesser extent Box-Ironbark. The remaining habitat types (Foothills woodland and Montane forest) could not be easily distinguished from the unclassified sites. If we had no evidence for prior groupings in these data, we might use a minimum spanning tree to further examine relative closeness of sites (Figure 18.5(b)). The three longest spans would roughly separate the River Red Gum and Gippsland Manna Gum habitats from the rest, with two of the unclassified sites intermediate.

Mac Nally (1996) was able to classify the sites *a priori* into five habitat types so we were able to test the $H_0$ of no difference between the five habitat types using a single factor ANOSIM procedure. We used the program PRIMER. The global *R* statistic was 0.914 and the probability of obtaining a value this great or greater, based on a randomization test, was less than 0.001. We concluded there were statistically significant differences in bird assemblages between habitats. Pairwise ANOSIM tests were difficult to interpret because there were only four observations in each group which only allowed 35 possible permutations for each pairwise randomization test and thus *P* values were ± approximately 0.029. However, only two of the pairwise comparisons had *R* values less than one, Montane forest versus Foothills woodland and Box-Ironbark versus Foothills woodland.

We also used the non-parametric MANOVA procedure of Anderson (2001) to test the $H_0$ of no difference between the five habitat types. We used the program NP-MANOVA, kindly supplied by M.J. Anderson from the University of Auckland. The single factor MANOVA test was based on Bray–Curtis dissimilarities between sites and we used 10 000 permutations.



| Source | SS | df | MS | *F* | *P* | Possible number of permutations |
|---|---|---|---|---|---|---|
| Habitat | 28 964.903 | 4 | 7241.226 | 9.619 | <0.001 | $2.55 \times 10^9$ |
| Residual | 11 292.165 | 15 | 752.811 | | | |
| Total | 40 257.068 | 19 | | | | |

Clearly, we would reject the $H_0$ and conclude that there is a significant difference across the five habitats in the Bray–Curtis dissimilarities between sites. We then ran pairwise comparisons, based on *t* statistics ($\sqrt{F}$ from non-parametric MANOVA comparing two groups). All comparisons were significant, except Foothills woodland v Montane forest, indicating that this procedure is more powerful than the ANOSIM tests, although Holm's adjustment to the *P* values to control the family-wise Type I error rate resulted in no significant differences (all $P = 0.280$).

| Comparison | *t* | *P* |
|---|---|---|
| Box-Ironbark v Foothills woodland | 1.702 | 0.031 |
| Box-Ironbark v Gippsland Manna Gum | 3.227 | 0.028 |
| Box-Ironbark v Montane forest | 2.676 | 0.028 |
| Box-Ironbark v River Red Gum | 3.639 | 0.028 |
| Foothills woodland v Gippsland Manna Gum | 2.954 | 0.031 |
| Foothills woodland v Montane forest | 1.520 | 0.054 |
| Foothills woodland v River Red Gum | 3.550 | 0.028 |
| Gippsland Manna Gum v Montane forest | 3.262 | 0.029 |
| Gippsland Manna Gum v River Red Gum | 3.361 | 0.028 |
| Montane forest v River Red Gum | 4.287 | 0.030 |

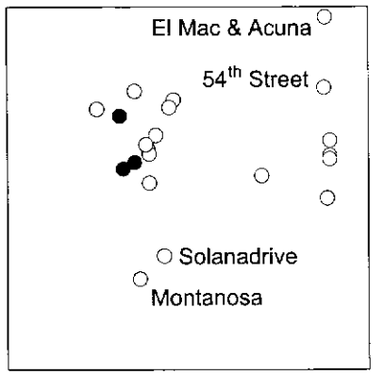



**Figure 18.4** NMDS scaling/ordination plot of the 28 sites from Bolger *et al.* (1997) based on a Bray–Curtis matrix of dissimilarities between sites, standardized so all sites have maximum abundance of one. The three mainland sites are filled symbols.

between minimizing stress and minimizing the number of dimensions. Our experience with ecological data is that two or three dimensions will usually produce adequate configurations.

The final orientation of the configuration is arbitrary and it is only the relative distances between objects that are relevant to interpretation in MDS. It is preferable to rotate the final configuration so that the first axis lies along the direction of maximum variation. This can be achieved by a PCA on the MDS axis scores (Clarke & Warwick 1994) and will often be done automatically by MDS software. Note that actual values of the object scores are also arbitrary and these can be scaled in a number of ways; only the relative distances between the objects is important. Plots of the final configuration do not need scales on the axes as long as the axes are scaled identically.

Basically, the interpretation of final scaling (ordination) plot is subjective. Objects closer together are more similar (e.g. in species composition) than those further apart. A useful addition to the plot is a minimum spanning tree, where the objects are joined by lines so that the sum of line lengths is the smallest possible and there are no closed loops (Figure 18.5(b)). Minimum spanning trees can be applied to any scatterplot of points. For MDS configurations, objects joined by the shortest spans are closest on the plot and those separated by longest spans are furthest

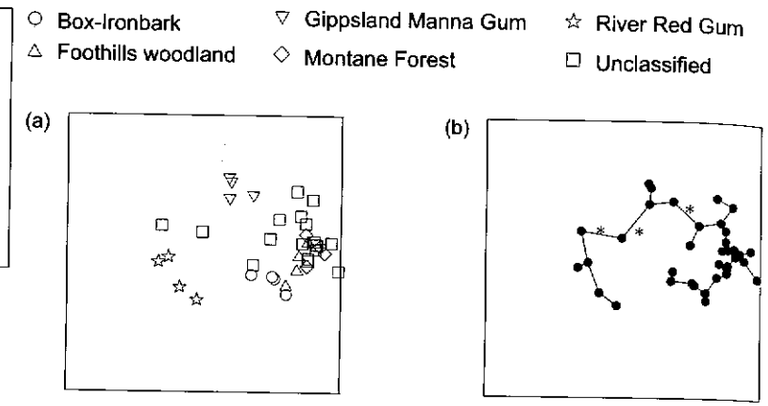



**Figure 18.5** NMDS scaling/ordination plots of the 37 sites from Mac Nally (1989) based on a Bray–Curtis matrix of dissimilarities between sites. In (a), the different habitats are identified by different symbols. In (b), a minimum spanning tree joins all sites with longest spans indicated by *.

apart; the latter may separate different groups of objects (see Digby & Kempton 1987). Minimum spanning trees can be plotted in three dimensions, although they become ugly to interpret.

We may also have formal hypotheses we wish to test. For example, are dissimilarities between objects related to other differences, such as geographic distances? If the data consist of replicate objects within pre-defined natural (e.g. polluted area vs non-polluted area) or experimental (e.g. different nutrient treatments) groups, then we would probably test whether objects within a group are closer together than objects from different groups. Testing these hypotheses will be considered in Sections 18.1.3 and 18.1.5.

**Convergence problems**

The algorithms for enhanced MDS converge to the final configuration iteratively and the number of iterations depends on the complexity of the data. More rapid convergence can be achieved if the coordinates from an initial PCA or PCoA scaling are used rather than a random starting configuration and some software for MDS defaults to a preliminary PCoA before iterating. The iterative nature of the various algorithms for enhanced MDS means that the iterations can converge to a "local" solution that is not the configuration that best matches inter-object distances with dissimilarities. The only solution to this problem is to repeat the MDS a number of times, using a new random starting configuration each time, and then compare the different configurations for stress and axis coordinates. We can only be confident of the final configuration if it occurs from a majority of random starts. Comparison of different configurations can be achieved through Procrustes analysis (Digby & Kempton 1987), where one configuration is rotated and rescaled to most closely match a second configuration of the same objects. The fit is measured by the sum of squared distances between the corresponding objects in the two configurations.

### 18.1.3 Dissimilarities and testing hypotheses about groups of objects

It is common for biologists to have recorded multiple variables from objects in a sampling or experimental design where the objects fall into pre-defined groups. The design might have a single factor or be multifactorial with factors either crossed or nested. We would often be interested in testing null hypotheses about differences between groups in these designs, as we would using linear ANOVA models if we had just a single response variable. In the multivariate context, the methods for testing such hypotheses proposed in the literature are based on the original variables, the scores for each object in scaling (ordination) space or the dissimilarities between objects.

Tests based on dissimilarities are not straightforward for two reasons. First, the dissimilarities between objects are not always independent of each other (the dissimilarities between objects 1 and 2 and 2 and 3 are not independent of the dissimilarity between objects 1 and 3), so randomization (permutation) testing procedures are required (Chapter 3). Second, if we wish to use the dissimilarities in linear models, we require sums-of-squares based on the difference between each



observation and the mean of the observations, or the centroid in the multivariate context. When dealing with metric dissimilarities (e.g. Euclidean distance), the centroid of a group of observations and the sum of squared deviations from this centroid are straightforward to calculate and interpret. This is not the case when dealing with non-metric dissimilarities like Bray–Curtis and a limitation of some approaches is their inability to deal with non-metric dissimilarities (see Anderson 2001).

**MANOVA based on original variables**

We could use a multivariate analysis of variance (MANOVA; see Chapter 16), a multivariate analogue of the univariate ANOVA, to test the null hypothesis of no difference between groups in some linear combination of variables. While MANOVA may be useful in some situations, it has quite restrictive assumptions about variances and covariances that are difficult to test (Chapter 16) and are unlikely to be met when the variables are species abundances with lots of zeros. A robust non-parametric form of MANOVA (NPMANOVA) that uses dissimilarities has recently been described by Anderson (2001) and will be discussed below. MANOVA comparing groups of objects is also restricted to data sets where the number of variables does not greatly exceed the number of objects, whereas ecological data sets often comprise many variables (species) and fewer objects (sampling untis).

**(M)ANOVA based on axis scores**

Another approach is to use any of the scaling procedures from Chapter 17 or this chapter that provide scores for each object on derived variables (components or axes). These scores could be used as response variables in linear models, as described for PCA in Chapter 17, to test hypotheses about group differences. There are some problems with this method. With MDS, we have to decide which axes to use; maybe scores from multiple axes (i.e. the first two or three dimensions if stress is adequate) could be used with a MANOVA? The axes themselves are also not a linear combination of variables like the components from a PCA or axes from a CA so are more difficult to relate to the original variables. Finally, the MDS axes simply define the relative positions of the objects in multidimensional space so as to represent the observed dissimilarities. Tests of hypotheses about group differences might be better based on these actual dissimilarities rather than some approximation of them.

**Mantel test**

The Mantel test described in Chapter 15 can be used to correlate a dissimilarity matrix between objects with another dissimilarity matrix that simply separates objects into groups (Manly 1997, Schnell *et al.* 1985). This second matrix is termed the model or design matrix (Legendre & Legendre 1998, Sokal & Rohlf 1995). The main limitation of using the Mantel test in this way is that it is difficult to test more complex models such as those including interaction terms.

Rundle & Jackson (1996) used a Mantel test to test for differences in the fish communities of the littoral zones of three lakes in Canada based on five sites in each lake. They constructed a Bray–Curtis dissimilarity matrix between the 15 sites. To test whether the variation in fish communities was primarily between lakes rather than within lakes, they used Mantel test to assess whether the Bray–Curtis matrix based on fish was associated with a matrix containing zeros for within-lake distances between sites and ones for between-lake distances between sites.

**Multi-response permutation procedures**

Mielke *et al.* (1976, see also Mielke 1985) proposed multi-response permutation procedures (MRPP) that test hypotheses about group differences in Euclidean distances, and Zimmerman *et al.* (1985) illustrated their application to biological data sets, such as $n$ sampling units by $p$ species. Basically, the MRPP determines the mean of the Euclidean distances between objects within each group and calculates an MRPP statistic (delta) that is a linear combination of these mean within-group Euclidean distances. The statistic produces a weighted average (based on sample size) of the within-group mean Euclidean distances. Small values of the statistic indicate that objects tend to be found in groups. The probability distribution of the MRPP statistic is determined by randomizing the allocation of all objects to the groups,

keeping the original sample sizes, with the null hypothesis being that all random allocations are equally likely. We compare our observed value of the MRPP statistic to the probability distribution generated under randomization to get the probability of obtaining the observed value of the statistic or one smaller under the null hypothesis. The MRPP can be used for a range of hypotheses including those associated with paired comparisons and randomized block designs.

MRPPs have been traditionally based on Euclidean distance and their use with more robust non-metric dissimilarities would be tricky because of the difficulty of defining the centroid and calculating the mean within-group dissimilarity. Nonetheless, McCune & Mefford (1999) have suggested that MRPPs might work well with other dissimilarity measures, such as Bray–Curtis. Since Euclidan distance is not a particularly appropriate measure of dissimilarity for some types of biological data, e.g. species abundances (Chapter 15), we could use the inter-object distances from classical (PCoA) or enhanced scaling (NMDS) in a MRPP. This is not an ideal solution because we know that these distances are an imperfect representation of the actual dissimilarities, and correction for negative eigenvalues would be required for PCoA. This approach is used, although not for MRPP, in distance-based redundancy analysis (Legendre & Anderson 1999a) and discussed below.

### Analysis of similarities

ANOSIM (Analysis of Similarities; Clarke 1993, Clarke & Warwick 1994) is a hypothesis testing procedure that uses Bray–Curtis dissimilarities, although it could use any dissimilarity measure. This procedure uses a test statistic ($R$) based on the difference between the average of all the rank dissimilarities between objects between groups ($\bar{r}_B$) and the average of all the rank dissimilarities between objects within groups ($\bar{r}_W$):

$$R = \frac{\bar{r}_B - \bar{r}_W}{n(n-1)/4} \qquad (18.2)$$

This is analogous to an ANOVA comparing between-group and within-group variation. The use of rank dissimilarities rather than actual dissimilarities is in keeping with the spirit of non-metric MDS.

The $H_0$ being tested by ANOSIM is that the average of the rank dissimilarities between all possible pairs of objects in different groups is the same as the average of the rank dissimilarities between pairs of objects in the same groups. $R$ is scaled to be within the range +1 to −1. Differences between groups would be suggested by $R$ values greater than zero where objects are more dissimilar between groups than within groups. $R$ values of zero indicate that the null hypothesis is true. Negative $R$ values indicate that dissimilarities within groups are greater than dissimilarities between groups, an outcome Clarke & Warwick (1994) considered unlikely. However, Chapman & Underwood (1999) showed that negative $R$ values can occur, especially when groups had high levels of within-group variability that were similar between groups and when outliers were present. They argued that negative $R$ values could be a useful diagnostic, indicating an inappropriate completely random sampling design when stratified sampling would be more appropriate.

Like the MRPP, ANOSIM uses a randomization procedure to randomly allocate objects to groups to generate the distribution of $R$ under the null hypothesis that all random allocations are equally likely. Clarke & Warwick (1994) described the use of ANOSIM procedures for nested designs where averaging over the subsampling levels produces a series of single factor tests for each factor. They also proposed ANOSIM for testing main effects in factorial designs by simply treating each main effect as a single factor test, averaging over the other factor. Legendre & Legendre (1998) pointed out that ANOSIM is very similar to a Mantel test using a model matrix to define the groups specified in the hypothesis and the two methods should produce similar $P$ values for the same hypothesis.

Both MRPP and ANOSIM use some measure of average dissimilarity within and between groups. Van Sickle (1997) described a useful graphical display for representing the relative strength of the differences in dissimilarity between groups, called a mean similarity dendogram. In its simplest form, a mean similarity dendogram for two or more groups would have branches for each group originating at the between-group mean dissimilarity and the length of each branch representing the within-group mean dissimilarities. Alternatively, the origin of each group branch could be staggered, with the mean between-group dissimilarity for each pair of groups plotted separately. Displays for multifactor designs are also possible (Van Sickle 1997). Mean similarity dendograms use the actual mean dissimilarities, rather than their rank orders, for plotting and therefore do not provide a direct graphical representation of the ANOSIM results.

One of the limitations of both MRPP and the ANOSIM procedure is that complex tests, such as interaction terms in linear models, are not available. This is in part because tests of interactions are difficult in the randomization context, since the interaction hypothesis cannot be simply expressed in terms of a random reallocation of observations to groups (see slightly differing opinions in Edgington 1995 and Manly 1997). Interactions are most sensibly tested in a linear model framework that also considers main effects. Unfortunately, if non-metric dissimilarities like Bray–Curtis are used, it is not straightforward to partition the variance (sum-of-squares) from fitting a multivariate linear model because of the difficulty of defining deviations from the centroid of the observations (Anderson 2001, Legendre & Anderson 1999a).

### Distance-based redundancy analysis

Because of the difficulties in using MRPP or ANOSIM tests for designs with interactions, Legendre & Anderson (1999a; see 1999b for minor correction) proposed an alternative approach for testing group differences in dissimilarities, called distance-based redundancy analysis (db-RDA). Their method uses PCoA to convert the original dissimilarities into their equivalent Euclidean distances, correcting for negative eigenvalues (Section 18.1.1). The matrix of $n$ objects by $p$ principal coordinates is then related to grouping factors using redundancy analysis (RDA; Chapter 17), where the grouping factors are represented by a matrix of dummy variables (Chapter 5) and the relationship is tested by a linear model using randomization tests (Chapters 3 and 8). This makes it easy for testing interactions because the analysis just becomes a multiple linear regression model and any combination of crossed and nested, fixed and random factors can be included.

It turns out that we can get the same results by simply doing a MANOVA test on the corrected principal coordinates, although Legendre & Anderson (1999a) argued that db-RDA has the advantages of more robust randomization tests and does not require more objects than variables in the original data matrix. The latter advantage is important because ecological data sets nearly always have more species (variables) than sampling units (objects). The main limitation of db-RDA is its complexity and the need to have software for the RDA component.

### Non-parametric MANOVA

Distance-based RDA was developed to translate various non-metric measures of dissimilarity into their equivalent distance in Euclidean space using PCoA. We can then relate these distances to a design matrix using linear models (e.g. RDA) and calculate sum-of-squared deviations between observations and their centroid. McArdle & Anderson (2001) and Anderson (2001) have recently shown that the partitioning of sums-of-squares (SS) and variances used for testing linear models can also be applied directly to dissimilarities, even non-metric ones like Bray–Curtis. This method means that using PCoA on the original dissimilarities is not necessary and the negative eigenvalues produced by db-RDA correspond to negative SS. The correction for negative eigenvalues in db-RDA described by Legendre & Anderson (1999a) actually produces overly conservative tests when random factors are included in the design (McArdle & Anderson 2001).

The non-parametric MANOVA described by McArdle & Anderson (2001) and Anderson (2001) is elegantly simple and can be applied to any design structure. The main difficulty is developing a randomization test for complex terms like interactions (Chapter 9; see Manly 1997). Our view is that the non-parametric MANOVA is so widely applicable in the biological sciences that we will describe it in some detail.

Consider a single factor design with $p$ groups and $n$ objects in each group so the total number of objects is $N = pn$. For the equations below, any two objects are termed $h$ ($h = 1$ to $N$) and $i$ ($i = 1$ to $N$).

From an $N$ by $N$ matrix of dissimilarities ($d_{hi}$ e.g. Bray–Curtis) between all pairs of objects, we calculate three SS.

The first is the sum of squared dissimilarities between all pairs of objects divided by $N$:

$$SS_{Total} = \frac{1}{N}\sum_{h=1}^{N-1}\sum_{i=h+1}^{N} d_{hi}^2 \quad (18.3)$$

Note that only the lower (or upper) diagonal of the dissimilarity matrix is used. The dissimilarity between objects $h$ and $i$ is the same as between $i$ and $h$ and is only counted once in the calculation of $SS_{Total}$.

The second is the within-groups SS. The $SS_{Residual}$ is the sum of squared dissimilarities between objects within each group, summed over the groups:

$$SS_{Residual} = \frac{1}{N}\sum_{h=1}^{N-1}\sum_{i=h+1}^{N} d_{hi}^2 e_{hi} \quad (18.4)$$

In Equation 18.4, $e_{hi}$ equals one if object $h$ and $i$ are in the same group and zero if they are in different groups (just like the design matrix in the Mantel test above).

The between-groups SS is determined from the usual additive partitioning of the total SS described for ANOVA models in Chapter 8:

$$SS_{Groups} = SS_{Total} - SS_{Residual} \quad (18.5)$$

The approximate $F$-ratio statistic for testing the $H_0$ that all allocations of objects, and therefore dissimilarities between objects, between groups are equally likely is:

$$F = \frac{SS_{Groups}/(p-1)}{SS_{Residual}/(N-p)} \quad (18.6)$$

This is analogous to the $F$-ratio statistic for a single factor ANOVA model. The randomization test is then done in the same manner as described for single factor ANOVA tests in Chapter 8, using a subset of all possible permutations for anything except very small $p$ and $n$.

Pairwise contrasts of specific groups, either planned or unplanned, can be done using the same test statistic. If there are many contrasts, the significance levels may need to be adjusted to control family-wise Type I error rate, using one of the Bonferroni corrections described in Chapter 3.

However, the main advantage of this non-parametric MANOVA is that it can handle more complex designs, especially those that include interactions. Anderson (2001) provides appropriate formulae for factorial designs but the logic is straightforward. The $SS_{Total}$ are calculated using Equation 18.3. The main change from a single factor design is that we need to calculate within-groups SS for each factor separately, ignoring the other factor. The SS for each main effect are simply the difference between the $SS_{Total}$ and within-groups SS for that factor. The $SS_{Residual}$ are calculated using Equation 18.4 except that each combination of the two factors (each cell) is considered a single group. So the $e_{hi}$ equals one if the objects are in the same cell (combination of factors) and zero if they are in different cells. The $SS_{Interaction}$ are what is left after the main effects and residual SS are subtracted from the total. The F-ratios are determined following Equation 18.6, although the denominator may need to be changed if either factor is random (see Chapter 9).

As we discussed in Chapter 9, there are different approaches to randomization tests in factorial designs and some debate about whether randomization tests for interaction terms are possible. Manly (1997) summarized these different approaches, including whether to randomize observations or residuals and whether to impose restrictions on which objects are randomized for tests of different terms. He argued that the different methods produced comparable results.

We illustrate the use of a single factor non-parametric MANOVA with the bird community data from Mac Nally (1989) – see Box 18.4. There were four replicate sites for each of five forest habitats types; unclassified sites were not included in the comparison. There was a significant difference between habitats, although like the ANOSIM procedure, the small number of possible permutations with only four replicates per group meant that pairwise comparisons were difficult to interpret after adjusting significance levels. Based on raw $P$ values, the non-parametric MANOVA procedure seemed more powerful than the ANOSIM comparisons.

The two main advantages of the non-parametric MANOVA introduced by McArdle & Anderson (2001) and Anderson (2001) are that any dissimilarity measure can be used and the tests are based on the partitioning of sums-of-squares as used in classical linear models. This means that the method can be used for any design structure that can be formulated as a linear model (see Chapters 5, 6, 8–12) and can accommodate fixed and random factors by using different denominators in the approximate F-ratios. The only limitation is the difficulty of determining the appropriate randomization test procedure for complex designs.



### 18.1.4 Relating MDS to original variables

Another question of interest in scaling (ordination) procedures is to determine which variables contribute most to the observed pattern among objects, e.g. which species contribute most to the separation among sampling units or which morphological variables contribute most to the separation of organisms. As described in Section 18.1.3, we will often be using a sampling or experimental design that includes groups of objects and our interest will be which variables contribute most to the any separation among groups. When we scale using one of the $R$-mode methods described in the previous chapter, then we obtain loadings for each variable on each derived component (axis of the scaling plot) as in PCA or can plot object and variable scores jointly to examine correlations as in CA.

Scaling techniques that are based directly on dissimilarities, such as MDS, do not provide correlations between derived axis scores and variables as part of the algorithm but there are alternative ways of investigating how the variables contribute to the final configuration of objects. We could simply correlate the axis scores from an MDS with each variable or linear combination of variables. This is not an ideal solution because, besides the problem of increasing Type I error rates from multiple testing if we do numerous correlations, we have to decide how many and which dimensions from the MDS we use. Additionally, we know that the scores, or at least the distances between objects, are imperfect representations of the actual dissimilarities so a method that uses these dissimilarities directly would be preferable.

Clarke & Warwick (1994) described a procedure for ecological data termed SIMPER (similarity percentages) for determining which species (variables) are contributing most to the dissimilarity between groups of object (sampling units). For example, the Bray–Curtis dissimilarity for a pair of sampling units is basically the differences between the units for each species, summed over all the species. SIMPER computes the percentage contribution of each species to the dissimilarities between all pairs of sampling units in different groups and the percentage contribution of each species to the similarities between all pairs of sampling units within each group. It then calculates the average of these percentage contributions, with its standard deviation. Species with a large ratio of average/standard deviation percentage contribution to dissimilarity between sampling units in different groups are those species that best discriminate between the groups. Note that there are no formal tests of hypotheses with SIMPER, just a list of species in order of their percentage contributions to dissimilarities between groups or similarities within groups.

### 18.1.5 Relating MDS to covariates

In ecological data sets, we often have two types of variable recorded for each sampling unit, species abundances (or presence/absence) and environmental characteristics (covariates). In these circumstances, we might wish to relate the dissimilarities between sampling units, or groups of sampling units, based on the species variables to differences in the environmental characteristics. Are sampling units that are very different from others in terms of species composition also very different in terms of one or more environmental variables? There are numerous ways of relating dissimilarities between sampling units to environmental variables, two of which we have already described. We could examine correlations between, or fit regression models to, the scores for each axis from the MDS and the environmental variable(s) (Ludwig & Reynolds 1988), just as we described for component scores from a PCA in Chapter 17. These correlations can be represented as vectors on the MDS plot, producing a biplot, and tests of the correlations are best done in a randomization context. The problems with relating environmental variables (covariates) to axis scores are the same as outlined in Sections 18.1.3 and

18.1.4, i.e. the problem of multiple testing, axis scores being an imperfect representation of the actual dissimilarities, deciding how many and which dimensions to use.

Clarke & Ainsworth (1993) proposed a procedure for ecological data that basically measures the correlation between dissimilarities between sampling units based on species composition and the dissimilarities between sampling units based on environmental variables. They provided an algorithm called BIO-ENV that first calculates a dissimilarity matrix (e.g. Bray–Curtis) between sampling units based on species abundances and a separate dissimilarity matrix (e.g. Euclidean distance) between sampling units based on environmental variables. It then measures any correlation between the rank-orders of these two matrices using the Spearman rank correlation coefficient. Each pair of observations for the correlation will be the rank of the Bray–Curtis dissimilarity (from species abundances) between objects $h$ and $i$ and the rank of the Euclidean distance (from environmental variables) between objects $h$ and $i$.

Legendre & Legendre (1998) pointed out that the BIO-ENV procedure basically calculates the same correlation as a Mantel test (Chapter 15 and Section 18.1.3), except the former is based on rank transformed data. The Mantel test could be used for the global test of no correlation between the two matrices, or even between the dissimilarities based on species composition and differences between sampling units for each environmental variable separately. It can also be extended to compare more than two matrices (Diniz-Filho & Bini 1996).

Clarke & Ainsworth (1993) and Clarke & Warwick (1994) incorporated a stepwise routine into their BIO-ENV procedure, to find the combinations of environmental variables that produce dissimilarities between sampling units with the highest correlations with dissimilarities between sampling units based on species composition. They argued that their implementation of the Mantel test is not suitable for hypothesis testing, both because the dissmilarities for both sets of variables are not independent and also because their stepwise procedure would produce numerous significance tests that are difficult to interpret (see Chapter 6).

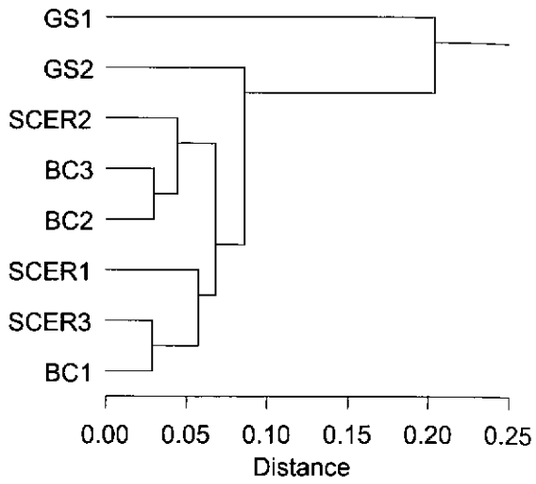



**Figure 18.6** Dendrogram from hierarchical UPGMA cluster analysis of the eight subpopulations of the plant *Clarkia springvillensis* based on Cavalli–Sforza genetic distances between subpopulations; from McCue *et al.* (1996).

Procrustes analysis (Section 18.1.2; Digby & Kempton 1987, Legendre & Legendre 1998) can also provide a descriptive measure of the fit of a configuration between objects based on one set of variables (e.g. species abundances) and a configuration between the same objects based on a separate set of variables (e.g. environmental characteristics).

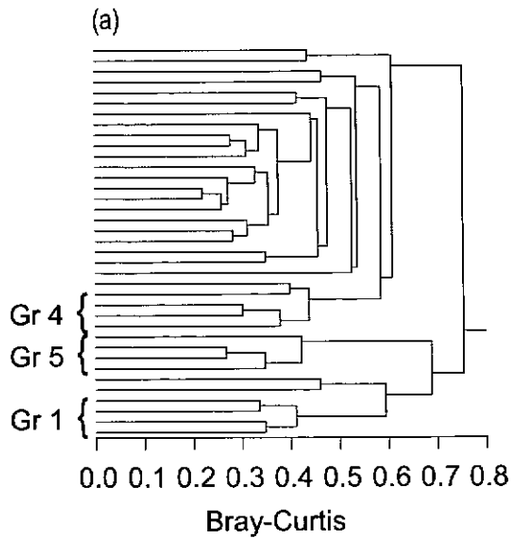
## 18.2 Classification

The aim of classification is to group together a number of objects based on their attributes or variables to produce groups of objects where each object within a group is more similar to other objects in that group than to objects in other groups. One form of classification analysis is discriminant function analysis (DFA; Chapter 16) where the number of groups was known *a priori*. In this section, we are interested in classification methods where the number of groups is not known and must be determined from the data.

### 18.2.1 Cluster analysis

Cluster analysis is a method for combining similar objects into groups or clusters, which can usually be displayed in a tree-like diagram, called a dendrogram (Figure 18.6). Legendre & Legendre (1998) provide a recent, very thorough, discussion. Cluster analyses are used commonly by biologists. For example, Crews *et al.* (1995) examined plant species in montane rainforest in Hawaii. They compared six sites (varying in age) using the cover–abundance measures for numerous plant species. The objects were sites, the variables were species abundances and cluster analysis was used to place the sites into like groups. Koenig *et al.* (1994) studied acorn production in oak trees in California. They clustered five species of oaks (objects) based on twelve mean annual values of acorn production (variables). Probably the most important use of cluster analysis in biology is taxonomic and phylogenetic research, where the dissimilarity measures are often morphological or genetic/molecular differences between organisms, species, etc., and the dendrogram represents a possible evolutionary sequence.





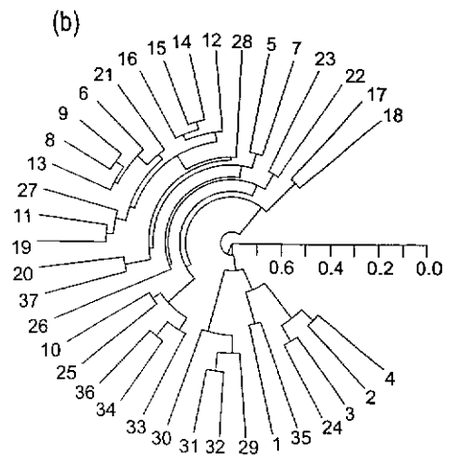




**Figure 18.7** Dendrograms from hierarchical UPGMA cluster analysis of the 37 sites from Mac Nally (1995) based on a Bray–Curtis matrix of dissimilarities between sites. In (a), the usual dendrogram is displayed with clusters identified for Gippsland Manna Gum (Gr 1), Box-Ironbark (Gr 4) and River Red Gum (Gr 5). In (b), the polar representation of the dendrogram is displayed, with site numbers. Gippsland Manna Gum includes sites 2, 3, 4 and 24; Montane forest sites 9, 11, 12, 15; Foothills woodland sites 10, 20, 21, 37; Box-Ironbark sites 25, 33, 34, 36; River Red Gum sites 29,30, 31, 32; remaining sites unclassified.

#### Agglomerative hierarchical clustering

Agglomerative methods start with individual objects and join objects and then objects and groups together until all the objects are in one big group. This is the form of cluster analysis familiar to most biologists. Usually objects are clustered but sometimes you may wish to cluster variables (e.g. species). Most algorithms for agglomerative cluster analysis start with a matrix of pairwise similarities or dissimilarities between the objects and the steps are as follows.

1. Calculate a matrix of dissimilarities ($d_{hi}$) between all pairs of objects.
2. The first cluster is formed between the two objects with the smallest dissimilarity.
3. The dissimilarities between this cluster and the remaining objects are then recalculated.
4. A second cluster is formed between cluster 1 and the object most similar to cluster 1.
5. The procedure continues until all objects are linked in clusters.

The graphical representation of the cluster analysis is a dendrogram (Figure 18.6, Figure 18.7(a)), showing the links between groups of objects with the lengths of the lines representing dissimilarity. If there are many objects, the standard dendrogram can be very long and difficult to represent on a single page. An alternative representation is the polar dendrogram (Figure 18.7(b)),

## Box 18.5 | Worked example of cluster analysis: genetic structure of a rare plant

Like MDS, hierarchical cluster analysis can be based on any type of dissimilarity matrix. We clustered the data on the eight subpopulations of the rare annual plant (*Clarkia springvillensis*) in California based on Cavalli–Sforza genetic distances between the subpopulations (McCue *et al.* 1996). We used UPGMA and the dendrogram is shown in Figure 18.6. The two Gauging Station subpopulations (GS) split off first; these were most different in the NMDS scaling plot based on the same matrix – see Box 18.2. Then the second of the Springville Clarkia Ecological Reserve (SCER) subpopulations grouped with the second and third Bear Creek (BC) subpopulations and the first BC subpopulation grouped with the first and third SCER subpopulations.

where the objects are arranged in a circle and their distance from the center of the circle represents dissimilarities between objects and groups of objects. Like scaling (ordination) plots, the interpretation of the groupings in the dendrogram is subjective and the decision about which groups to report is usually based on some arbitrary cut-off value for dissimilarity.

The major difference between the variety of available hierarchical agglomerative clustering methods is how the dissimilarities between clusters and between clusters and objects (step 3) are recalculated. These are termed linkage methods, and three common ones are as follows.

- Single linkage (nearest neighbour), where the dissimilarity between two clusters is measured by the minimum dissimilarity between all combinations of two objects, one from each cluster.
- Complete linkage (furthest neighbour), where the dissimilarity between two clusters is measured by the maximum dissimilarity between all combinations of two objects, one from each cluster.
- Average linkage (group average or mean), where the dissimilarity between two clusters is measured by the average of all the dissimilarities between all combinations of two objects, one from each cluster. The group mean (or average) linkage strategy, commonly called unweighted pair-groups method using arithmetic averages (UPGMA), is often recommended. There is a weighted version of UPGMA (WPGMA), which weights the original dissimilarities differently, and unweighted clustering based on centroids (UPGMC), which is equivalent to UPGMA except that centroids instead of means are used.

Kent & Coker (1992), Legendre & Legendre (1998) and Ludwig & Reynolds (1988) discuss the pros and cons of these different linkage methods. If there are "strong" (i.e. very dissimilar) groups in your data, then the different methods will produce similar dendrograms; in contrast, the different linkage strategies can produce very different patterns for data with weak structure (Ludwig & Reynolds 1988). Belbin *et al.* (1993) proposed a flexible modification of UPGMA that allowed the clusters to be better, if artificially, defined and this method effectively recovered true groups in the data based on simulation studies (Belbin & McDonald 1993).

Box 18.5 illustrates a cluster analysis of the subpopulations of *Clarkia springvillensis* based on genetic differences recorded by McCue *et al.* (1996). A cluster analysis of the 37 sites in southeastern Australia, using Bray–Curtis dissimilarities based on the densities of 102 species of forest birds (Mac Nally 1989), is presented in Box 18.6.

Agglomerative cluster analysis does have some disadvantages, primarily related to the interpretation of the dendrogram. The hierarchical approach means that once a group or cluster is formed from two or more objects, that group cannot be broken later in the process. As a result, the dendrogram is not a representation of all pairwise dissimilarities between objects like in multidimensional scaling (MDS). A misleading cluster formed early in the process will influence the remaining clusters. Also, the analysis forces objects into clusters and it would be easy for naïve biologists to place too much emphasis on these clusters without examining the actual dissimilarities. We much prefer MDS as a method for graphically representing relationships between objects based on dissimilarities.



## Box 18.6 | Worked example of cluster analysis: geographic variation and forest bird assemblages

A matrix of Bray–Curtis dissimilarities, based on densities of 102 species of birds, between sites was used to hierarchically structure the 37 sites in southeastern Australia (Mac Nally 1989). No standardization was used because the data were densities of birds, rather than absolute counts. The UPGMA clustering procedure produced the dendrogram shown in Figure 18.7(a), although representing this in polar form (Figure 18.7(b)) makes presentation a little easier. The Gippsland Manna Gum sites, the River Red Gum sites and the Box-Ironbark sites grouped into clear clusters, whereas the remaining habitat types (Foothills woodland and Montane forest) were not in separate clusters. This interpretation is similar to that from the NMDS on the same matrix of dissimilarities (Box 18.5).

### Divisive hierarchical clustering

Divisive methods have a long history for clustering ecological data. They basically start with the objects in a single group and split them up into smaller and smaller groups until each group is a single object. One method popular with ecologists is two-way indicator species analysis (TWINSPAN), a complex procedure that uses the reciprocal averaging algorithm of correspondence analysis (Chapter 17) to successively divide the first axis for both sampling units and species into smaller groups. The output includes a two-way table that orders the sampling units and species and shows the groupings and the relative abundances of species for each sampling unit. The actual computations are tedious, although a detailed description can be found in Kent & Coker (1992). Van Groenewood (1992) and Belbin & McDonald (1993) provided simulation results that showed that TWINSPAN is not particularly good at detecting true clusters in ecological data and the problems that affect correspondence analysis, particularly the distortion of sampling units along the first axis, also affect TWINSPAN.

### Non-hierarchical clustering

Non-hierarchical methods do not represent the relationship between objects in hierarchical form. Basically, they start with a single object and cluster other objects that are similar to the first one. In contrast to hierarchical clustering, objects can be reassigned to clusters during the clustering process. One method common in statistical software is *K*-means clustering – see Legendre & Legendre (1998) for a detailed description. *K*-means works by splitting the objects into a predefined number (*K*) of clusters, and then cluster membership of objects is iteratively re-evaluated by some criterion, such as to maximize the ratio of between-cluster to within-cluster variance. Another method is additive tree clustering, which develops a tree-like network (dendrogram) where the dissimilarity between objects within a cluster is represented by the sum of the lengths of the branches joining them (Gower 1996) and may be more suited to non-metric dissimilarity measures.

## 18.3 | Scaling (ordination) and clustering for biological data

When the main purpose of the multivariate analysis is to scale objects, what ecologists term ordination, numerous techniques are available. There

have been many evaluations and comparisons of these techniques, particularly for ecological data in the form of species abundances across sampling units. Differing opinions on the relative merits of different techniques can be found in Faith *et al.* (1987), Jackson & Somers (1991), Minchin (1987), Palmer (1993), Peet *et al.* (1988), ter Braak & Verdonschot (1995), van Groenewood (1992), and Wartenberg *et al.* (1987), among others. In our view, the choice of method depends on the nature of the data, the implicit measure of dissimilarity used by each method, and, not surprisingly, the biological question being addressed. Our preferred approach is to use a method that is applicable to a range of data types, is amenable to various user-defined standardizations and transformations of the data, is flexible in terms of which dissimilarity measure is used, and can be used for describing patterns and testing *a priori* hypotheses. Multidimensional scaling (MDS), especially the robust non-metric version (NMDS), meets all these criteria. Any measure of dissimilarity can be used, thereby allowing dissimilarities between objects based on continuous, binary and mixed variables under nearly every combination of transformation and standardization. The scaling or ordination has been shown to be robust for a range of data types, accurately representing underlying true dissimilarities and recovering ecological gradients, and hypothesis tests can be based on the dissimilarities. For ecological data, NMDS also appears to be the most robust for nonlinear relationships of species abundances across sampling units along long ecological gradients, which can result in misleading arching of second and higher dimensions in some methods.

The most obvious competing technique is correspondence analysis (CA) or the more sophisticated canonical version (CCA). The strengths of these methods are also their weaknesses. By implicitly using the chi-square metric as the dissimilarity measure, they allow joint scaling plots of objects and variables and when axes are scaled similarly, relative positions of objects and variables can be compared. Unfortunately, the restriction to the chi-square metric also reduces flexibility and this dissimilarity measure may not be ideal for some forms of data (Faith *et al.* 1987). There are also decisions to be made about how to scale the axis scores, although the different scalings don't often alter the general pattern from the joint plot.

Constrained ordinations like CCA and redundancy analysis (RDA) also allow for biplots, where covariates can be included on the scaling plot showing which axes are correlated with which covariates. This is probably the main reason for the popularity of these methods, especially CCA. Relationships between dissimilarities and covariates under the MDS framework can also be evaluated although not in the same direct manner as in CCA and RDA. Finally, we shouldn't forget the oldest of these techniques, principal components analysis (PCA). While not always suitable as a scaling/ordination procedure, PCA is still a very important method for variable reduction, especially when linear relationships between variables are expected.

You may have inferred from Section 18.2.1 that we are not big users of cluster analysis, especially for representing dissimilarities between objects. Clustering procedures do not really use all pairwise dissimilarities for grouping objects so the dendrogram is not necessarily a good representation of a dissimilarity matrix. The main use of clustering procedures in biology is to display possible evolutionary and phylogenetic relationships, where the objects are organisms or taxonomic groups and the dissimilarities are morphological or genetic differences. Cluster analysis has less applicability for analyzing species abundance data to show relationships among sampling units. Ecologists sometimes use an initial cluster analysis to identify groups in a data set and then indicate those groups on a subsequent scaling plot. This approach has never made much sense to us, the cluster analysis almost certainly being a less efficient way of representing dissimilarities between objects than a method like enhanced MDS (but see Legendre & Legendre 1998 for an alternative view). Certainly, it is inappropriate to test hypotheses about differences between these groups; hypothesis tests cannot be validly used to compare groups that were defined by the same data.



## 18.4 General issues and hints for analysis

### 18.4.1 General issues

- Principal coordinates analysis (PCoA) is a useful metric scaling procedure but has generally been superseded by enhanced, iterative scaling procedures.
- Our preferred technique for scaling or ordination of ecological data, when there are numerous zeros and extracting underlying ecological gradients is important, is a combination of a suitable dissimilarity measure, like Bray–Curtis, and robust non-metric multidimensional scaling.
- Non-metric MDS is probably more robust than metric MDS, especially when the relationship between dissimilarities and inter-object distances is nonlinear. Hybrid MDS may offer a slight advantage.
- Hierarchical cluster analysis is not as useful as MDS for representing a dissimilarity matrix and has the disadvantage of forcing all objects into clusters that cannot be reassessed during the clustering procedure.

### 18.4.2 Hints for analysis

- Final enhanced MDS configurations should not be interpreted without examining stress values. Make sure you know which version of stress your software uses. Values for version one of Kruskal's stress should be less than 0.15, ideally less than 0.10, for configurations of objects to be considered reliable.
- Multiple runs from random starting configurations should be compared with enhanced MDS, to ensure that any configuration does not represent a local, unrepeatable, pattern. With large data sets, i.e. many objects, using an initial PCoA to determine a starting configuration may help convergence.
- Analysis of similarities (ANOSIM) or multiresponse permutation procedures (MRPP) are useful ways of testing hypotheses about group differences in a multivariate context, the former retaining the underlying philosophy of NMDS. For pairwise comparisons of groups, $n$ greater than four per group is needed for the randomization tests. For more complex hypotheses, especially tests of interactions, the non-parametric MANOVA of Anderson (2001) offers great promise.
- The unweighted pair-groups method using arithmetic averages (UPGMA) is usually recommended as a linkage strategy for agglomerative clustering. Non-hierarchical methods may offer more flexibility because clusters are not fixed once formed.

Chapter 19

# Presentation of results

A central part of reporting any scientific work is the presentation of the results, in either tabular or, more commonly, graphical form, and a considerable literature has accumulated about the appropriate ways for displaying quantitative data (e.g. Cleveland 1993, Tufte 1983, 1990, and some recent issues of *The American Statistician*). Much of this literature focuses on clarity of graphs and it is an issue that has become increasingly important as biologists do multifactorial experiments, often with complex underlying statistical models. We then face the problem of explaining those complex results to an audience that is pressed for time, and deluged by the number of papers published in any given month. In this environment, the presentation of your results becomes almost as important as the work itself, as you must convince a reader that he or she should persist with reading your paper, in the face of the many other demands on their time.

In many cases, the decision whether to read a paper completely is based initially on the title and abstract, which are provided by many of the electronic databases and the web. Having decided to look more closely at the paper, the next decision made is whether to persist with reading it. That decision will be made based in part on how clearly you express your ideas, and there is a long tradition of convincing scientists to write clearly, with several excellent and essential guides (e.g. Pechenik 2001, Strunk & White 1979, Williams 1997).

As a result of these issues, many of us think carefully about our writing style. In contrast, there is not such a long history of thinking about how to present the data, although there are some examples of creative ways to present the raw data from a study in a very complex appendix. Because the data and accompanying analyses determine whether the audience believes the story you are telling, it is critical that you present those results as clearly as possible, drawing attention to the most important features of the results, rather than submerging them in a sea of extraneous material. In this chapter, we present some simple ways to present analytical results and display results graphically, as well as making suggestions of ways not to present results. Our aim is not to be prescriptive about presentation, but to encourage you to think more about how to report your work.

## 19.1 Presentation of analyses

We will deal with some of the most common analyses, although many of these concerns and suggestions apply to a range of other statistical analyses.

### 19.1.1 Linear models

**Regression analyses**

Analyses of linear regression models are a clear example of where most statistics packages generate extensive output, but much of the information can be omitted. In the case of a simple linear regression with a single predictor variable, you will get an output similar to the one in Table 19.1 from most statistics packages.

The regression model examines the relationship between the number of limpets in a quadrat



**Table 19.1** Standard regression output from a major statistics package. The example is from SYSTAT version 6

```
Dep Var: LOGLIMP N: 40 Multiple R: 0.30.345 Squared multiple R: 0.119
Adjusted squared multiple R: 0.096
Standard error of estimate: 0.373

Effect    Coeff    SE     StdC  Tol  t       P
CONSTANT   1.072   0.083   0.0  .    12.994  0.000
ALGAE     -0.006   0.003  -0.3  1.0  -2.265  0.029

                    Analysis of Variance
Source      SS     DF  MS     F      P
Regression  0.713  1   0.713  5.129  0.029
Residual    5.282  38  0.139
```

(log-transformed, LOGLIMP) and the cover of algae (ALGAE), and the output gives us the estimated regression line, some measures of how precisely the parameters of the line – slope and intercept – have been estimated, and tests of hypotheses about the slope and intercept (by default, that each equals zero). Some or all of this material could be added into a table, but we can present most of the information in the text in standardized form.

First, in a simple regression, there is considerable redundancy. Most statistics packages are written to deal with complex regression models, and a simple regression is treated as just a special case of the general linear model. The bottom half of the output is an ANOVA table, testing whether the regression model (i.e. the set of predictor variables) explains significant amounts of the variation in the dependent variable. The top section of the table also shows tests of hypotheses – $t$ tests for the slope and intercept. With only one predictor variable, the ANOVA $F$ test and the $t$ test for the effect of algae are identical, and you can see on the output that the $F$-ratio of 5.129 is the square of $t$ (= 2.265), and the two $P$-values are identical (Chapter 5). There is, in this case, no point in reporting both values. Other parts of this output only become relevant when we have more predictor variables, e.g. tolerance, adjusted multiple $r^2$ (see Chapter 6). In most cases, we are interested only in whether the regression is significant, the estimates of the model parameters (which gives an idea whether the relationship is likely to be important), and some measure of how well the model fits the data. The $t$ or $F$ tests for the effects of the predictor variable provide the first information. The intercept and slope are listed under "Coeff" in the output table above ("CONSTANT" is often used to indicate the intercept of the regression model), and the simplest measure of the scatter of points around the line is the $r^2$, provided at the top of the output. We could therefore reduce that table of output to a single sentence in the text, using just the information highlighted on the output table:

> The number of limpets fell as algal cover increased, although algal cover only explained 12% of the variation in limpet abundance (equation: log(limpets) = 1.076 − 0.006 × algal cover, $F_{1,38} = 5.129$, $P = 0.029$, $r^2 = 0.119$).

This format is a standard one; and you could expect a reader to be familiar with the estimates of the parameters of the regression model, etc. – assuming that you've mentioned somewhere that it's a linear regression! If not, that information could be added inside the parentheses. Again, if we wished to be true minimalists or maximize the data density, we could omit the $r^2$ or even the $F$-ratio. As we discuss below if you know the df and

*P*, you can back-calculate the *F*-ratio, so the *P* and *F* are technically redundant. In the same fashion, for a simple regression, the $r^2$ can also be calculated from the ANOVA table – it's the $SS_{Regression}/(SS_{Regression}+SS_{Residual})$, so the *P*-value is enough for a desperate reader to calculate the $r^2$. We recommend this as overkill – most readers are comfortable with the information given in the previous paragraph. The only additional information might be interval estimates for the model parameters, such as confidence intervals.

The information from more complex regressions can also generally be compressed, although not to the same degree, and most complex regressions are presented in tables.

### ANOVA

The simplest way of presenting the results of a linear model with categorical predictor variables (i.e. a classical ANOVA model) is to display the complete ANOVA table. However, in many publication outlets, space is at a premium and there is usually pressure on authors of scientific papers in biology to reduce the amount of journal space devoted to results of statistical analyses. With this in mind, we should consider ways of presenting ANOVA results more efficiently without sacrificing information.

We suggest the following.

- The degrees of freedom should always be presented, as they indicate the sample size. Therefore, we do not need both SS and MS, as one can be calculated from the other using the df.
- As long as the $MS_{Residual}$ and *F*-ratios are provided, we don't need the MS for groups or specific contrasts as these can be calculated from the *F*-ratio, the degrees of freedom, and the $MS_{Residual}$. This step does require that you have described the statistical model adequately in the Methods section.
- As discussed in Chapter 3, we prefer *P*-values to be presented (at least for $P \geq 0.001$), as they allow readers to use their own significance level for testing $H_0$.

#### Single factor models

For a single factor ANOVA model, there is generally no need to report your findings in a table; there is only one way to calculate the *F*-ratio. You can report your analysis in the text of your results, giving results in a standardized form:

"Attending a stats course by the authors of this book did not markedly improve the quality of students' analyses ($F_{1,4} = 1.23$, $P = 0.546$)"

The information in parentheses tells a reader that the conclusion is based on an *F* test with numerator df = 1, denominator with df = 4, that the *F*-ratio is 1.23 (and hence the ratio of $MS_{Groups}$ to $MS_{Residual}$ is 1.23), and gives the probability of this value of *F*, or one larger, under the null hypothesis. There is no need for further information (except, perhaps, why this particular null hypothesis is retained). Of course, a real minimalist might argue that the value of the *F*-ratio is unnecessary; it follows automatically from the *P*-value and the two degrees of freedom.

If your analysis includes planned comparisons within an overall analysis, you can specify them in the same way, or you could include the analyses in a table. If you have listed the $df_{Residual}$ and $MS_{Residual}$ in the table, all you need to describe for most planned comparisons is the *P*-value. The vast majority of planned comparisons have numerator df = 1, and you have provided the other information (df, MS). In most cases, planned contrasts and trend analyses are best incorporated into the body of the ANOVA tables since they represent partitioning of the SS (see Chapter 8).

Multiple comparison results are commonly presented in two ways (i) labeling means in graphs and tables with the same letter or symbol if they are not significantly different, and (ii) listing the means (or group labels) in order and joining those not significantly different with an underline (see Chapter 8). The results can also be presented in the text, e.g. "a Tukey's test (with $\alpha = 0.05$) showed that the two highest densities had slower growth rates than the two lowest densities".

#### Complex ANOVA models

Two other issues are relevant for multifactor ANOVAs.

- When random factors are included, it is often useful to indicate the different error terms used in the ANOVA table unless it is a standard design. When the number of factors gets very large, there can be many possible (and actual) denominators, and a reader may not wish to derive the expected mean squares (e.g. Keough & Quinn 1998, for a very complex example).
- Measures of explained variance (e.g. variance components) are often incorporated into ANOVA tables when random factors are included.



### 19.1.2 Other analyses

Many other statistical methods also produce voluminous output with lots of redundancy, and you can generally reduce the volume of analytical results, without sacrificing information. There are also often conventions of how to report particular analyses – which pieces of information are critical to assure a reader that you know what you're talking about, and that you have results that he or she should believe. We will not go into details on these other analyses here, but, in the earlier chapters, you will find that the examples we cite can also be used as guides for how to report those kinds of analyses. Unfortunately, in some of the analyses that are only now making their way into the biological literature, such as randomization tests, including bootstraps and jackknifes, logistic regression, etc., there are no conventions for presenting analyses, and inspection of the literature shows great variation in how results are reported.

## 19.2 Layout of tables

Once you have decided which information to incorporate into a table, there is the matter of how the table can be laid out. Many current software packages allow you a wide range of formatting options, and, just like the discussion on graphic design in Section 19.3, some of those options improve the appearance of your text, while others produce hideous results. The table should be laid out to make the reader's job as easy as possible. Look at the examples in Box 19.1, and see which table provides the clearest layout of simple information. The tables present the results of testing for the effects of existing ascidians on settlement of marine invertebrates larvae. The analyses are single factor ANOVAs, and the table shows the *P* value from each analysis, together with an estimate of the residual variance and power values (to detect a 50% change in settlement rate). We had already decided to omit SS, MS, and F-ratios. The table is laid out simply, with no unusual formatting. The degrees of freedom were constant across species, and were detailed in the legend of the table.

We could improve the readability of the table by a few changes – there are three statistically significant results, and they can be highlighted by a bold typeface. The table shows results from two polychaetes, a species of barnacle, and a few bryozoans. If we want a reader to see them in their natural groups, and, perhaps, to contrast the results for different groups, we could either put faint lines between the groups or put some space between some of the lines. A reader then sees that all of the significant results fall in the same taxonomic group. In contrast, the lower panel shows one of the worst formatting styles, and we have buried the important information behind a large number of completely unnecessary grid lines. There is nothing to draw a reader's attention to the most important bits of information, in this case the tests of hypotheses, although we could equally have decided to highlight the power values.

In some cases, complex sets of results can best be displayed using non-standard table designs. For example, if there are many analyses of the same kind, such as analyses on a large number of species, the point of interest may be the patterns of significance, power, etc. For example, Table 19.2 shows an even simpler table, taken from Keough & Raimondi (1996), summarizing results from a whole suite of experiments. The experiments cover the effects of microbial films, and at issue is whether particular films stimulate, inhibit, or have no effect on settlement of larvae. In this table, the authors chose to use ticks and crosses to indicate positive and negative results and circles to indicate cases of no effect. A few weak or equivocal results are indicated by the "~". Blanks indicate that the species in question didn't settle during the particular experiment. Dotted lines separate groups belonging to different phyla. The table summarizes a large number of analyses from three papers.

These examples aren't an exhaustive list, nor are they necessarily the best ways to present information, but they do emphasize that there are alternatives to tedious standard ANOVA tables from complex models!

**Box 19.1** | Different arrangements of a table

| Taxon | *P* | $\sqrt{MS}_{Residual}$ | Power (ES = 50% ) |
|---|---|---|---|
| Serpulids | 0.348 | 20.32 | 100 |
| Spirorbids | 0.455 | 2.60 | 47 |
| *Elminius* | 0.531 | 24.89 | 71 |
| *Cryptosula* | 0.025 | 1.90 | 48 |
| *Scruparia* | 0.789 | 0.62 | 61 |
| *Tricellaria* | 0.017 | 4.72 | 98 |
| *Watersipora* | 0.525 | 3.45 | 94 |
| *Bugula neritina* | 0.118 | 10.36 | 69 |
| *Bugula stolonifera* | 0.042 | 18.60 | 100 |

| Taxon | *P* | $\sqrt{MS}_{Residual}$ | Power (ES = 50% ) |
|---|---|---|---|
| Serpulids | 0.348 | 20.32 | 100 |
| Spirorbids | 0.455 | 2.60 | 47 |
| *Elminius* | 0.531 | 24.89 | 71 |
| *Cryptosula* | **0.025** | 1.90 | 48 |
| *Scruparia* | 0.789 | 0.62 | 61 |
| *Tricellaria* | **0.017** | 4.72 | 98 |
| *Watersipora* | 0.525 | 3.45 | 94 |
| *Bugula neritina* | 0.118 | 10.36 | 69 |
| *Bugula stolonifera* | **0.042** | 18.60 | 100 |

| Taxon | *P* | $\sqrt{MS}_{Residual}$ | Power (ES = 50% ) |
|---|---|---|---|
| Serpulids | 0.348 | 20.32 | 100 |
| Spirorbids | 0.455 | 2.60 | 47 |
| *Elminius* | 0.531 | 24.89 | 71 |
| *Cryptosula* | 0.025 | 1.90 | 48 |
| *Scruparia* | 0.789 | 0.62 | 61 |
| *Tricellaria* | 0.017 | 4.72 | 98 |
| *Watersipora* | 0.525 | 3.45 | 94 |
| *Bugula neritina* | 0.118 | 10.36 | 69 |
| *Bugula stolonifera* | 0.042 | 18.60 | 100 |

## 19.3 | Displaying summaries of the data[1]

We will describe a number of different types of graphical display that biologists commonly use for summarizing numerical information and presenting results. In general, too little information is paid to the layout of these graphs, despite their being the sections of the paper that readers' attention is often drawn to first. There is a substantial literature on production of graphical display of information, including the excellent books by Tufte (1983, 1990), especially his wonderful 1983 book, and Cleveland (1994). The manual for the graphics component of the statistics package SYSTAT (SPSS 1999) includes an introductory chapter by Leland Wilkinson that is a clear discussion of graphic design.

1 In presenting the following, often hideous, graphs, you should be aware that we generally used the default settings of one or more common graphics packages, rather than trying hard to create awful graphs!



Table 19.2 | Layout of summary table highlighting results from several experiments

| | Variation in microbial cues | | | |
|---|---|---|---|---|
| **SETTLERS** | **presence absence** | **Short time** (0–6 d) | **Long time** (0–4 wk) | **Large spatial** (10s of km) |
| Serpulid polych. | ✓ | ✓ | ✓ | ✓ |
| Spirorbid polych. | ✓ | ✓ | ○ | ~ |
| *Elminius modestus* | ✘ | ✘ | ○ | ○ |
| *Balanus variegatus* | ~ | ✘ | ✓ | ○ |
| *Bugula neritina* | ~ | ~ | ✓ | ✓ |
| *Bugula dentata* | ✓ | ✓ | ✓ | ○ |
| *Bugula stolonifera* | ~ | ✓ | ✓ | ○ |
| *Tricellaria* | ~ | | ✓ | ✓ |
| Encrusting bryozoans | ✓ | ✓ | ✓ | ✓ |
| *Trididemnum* | ○ | | | ○ |
| *Botryllus schlosseri* | ~ | ○ | | ○ |
| *Didemnum* | ○ | ○ | ○ | ○ |
| *Diplosoma* | ○ | ○ | ○ | ○ |
| *Pyura stolonifera* | ○ | ○ | | |
| *Ascidia* | ○ | | | ○ |
| *Ciona intestinalis* | ○ | ○ | | ~ |
| Sponges | ✓ | ✓ | | ○ |
| *Electroma* | ○ | ○ | | |
| **Total recruitment** | ✓ | ✓ | ✓ | ✓ |

*Note:*
Ticks and crosses indicate positive and negative effects, circles for no effect and tildes for weak or equivocal results.

The guiding principle in constructing graphs is to produce clear, unambiguous, representations of your results. These representations should draw a reader's attention to what you consider the most important aspects of your results, and should be free of distracting elements. In most cases, this will mean simple, clean graphics, rather than the wonderfully ornate productions possible in many graphics packages. For complex experiments or sampling programs, this will entail decisions about which factors to include, which to highlight, etc.

Tufte coined phrases for some of what he saw as important problems.

- Data:ink ratios reflect the amount of ink need to present a given amount of data – high values are desirable.
- Data density is similar to the data:ink ratio, but reflects the space taken, rather than the ink used.
- Chartjunk is extraneous ornamentation that puts fancy things all around, but doesn't help explain your results. This a particular problem in many graphics packages used to prepare talks.

The way in which the information is presented will also vary with your target audience – a figure in a paper can be more complex than one that you might show at a conference, because the reader can sit and digest the information. Similarly, careful or thoughtful use of color can help an oral presentation, but most journals either don't permit colored graphs or impose an extra charge for colored figures. Newer electronic journals or

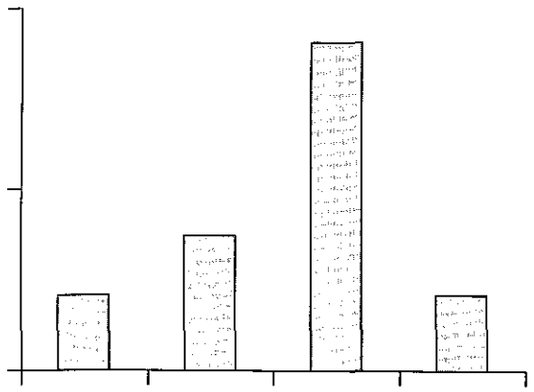


**Figure 19.1** Simple bar chart, showing means of four treatments.

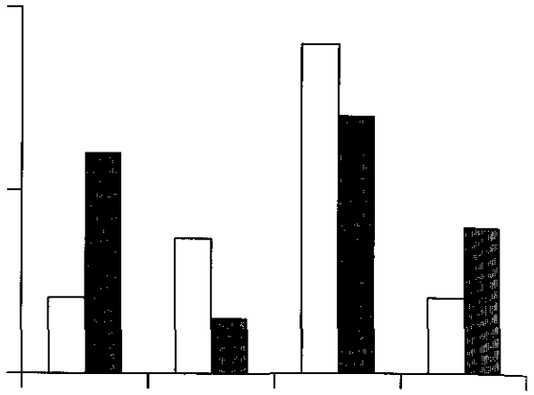


**Figure 19.2** Simple bar graph with two sets of four treatments, such as from a 4×2 factorial ANOVA.

similar outlets may permit color and reports, such as those from consultancies that involve a smaller number of copies, can use color to great advantage.

The technical limitations of the medium will also influence how you construct graphs. For example, in presenting a computer-based talk, you need to bear in mind that most computer projection facilities are still relatively low resolution (typically 640×480 or 800×600 pixels), and you can put less detail than on a high-resolution 35 mm slide. Similarly, many laser printers don't reproduce solid colors very well, and rather than solid black as a filling pattern, you may be better using hatched or stippled fill patterns. The same cross-hatched patterns may look awful as part of a computer presentation, when solid colors work much better. Tufte refers to some of the unfortunate choices of fill patterns as "unintentional optical art"!

In talking about graphics, we focus on the most common ways of displaying information.

### 19.3.1 Bar graph

A bar graph is used to plot some quantitative variable on the *Y*-axis against a grouping (categorical) variable on the *X*-axis, where the value of the variable for each category is represented by the height of a rectangular bar (Figure 19.1). The width of the bars can be altered to improve aesthetic appearance. The top of the bar may represent a single value or it may represent a summary statistic, such as a mean. In the latter case, some measure of variation or precision should be provided using error bars (Figure 19.14; also see discussion on error bars below).

If there is a second grouping variable, then it can be represented by adjacent bars, with different fill patterns or colors, at each level of the first grouping variable (Figure 19.2). A variation on bar graphs sometimes used in business presentations is called a pictogram (Snee & Pfeifer 1983), where the bar is replaced by objects which illustrate the variable being plotted, e.g. some product. We eschew pictograms because the actual value represented by the object is sometimes difficult to determine (see, for example, Chapter 2 of Tufte 1983), and there is no sensible way to include error bars.

Some attention should be paid to the fill patterns, too – as lamented by Tufte, most modern software packages give you access to a wide range of fill patterns, many of them appearing to have been designed while blindfolded. Fill patterns should not distract the reader – remember that particular kinds of hatching can cause adjacent bars to blur, or make it difficult to see where objects really end. For example, Figure 19.3 shows some samples of awful fill patterns or a poor choice of fill patterns to be alongside each other. Again, we've done nothing special here – these patterns are standard options of a common software package. On the left-hand panel, adjacent bars with poor cross-hatching make the information



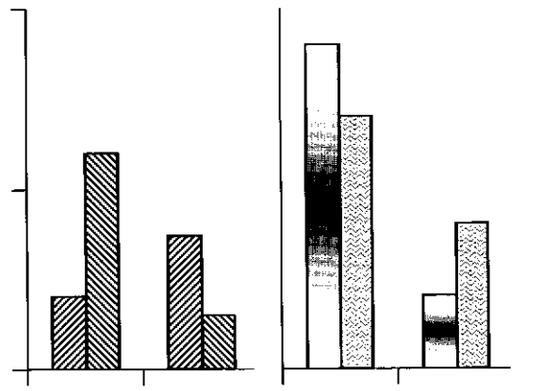


**Figure 19.3** Examples of fill patterns from standard software packages.

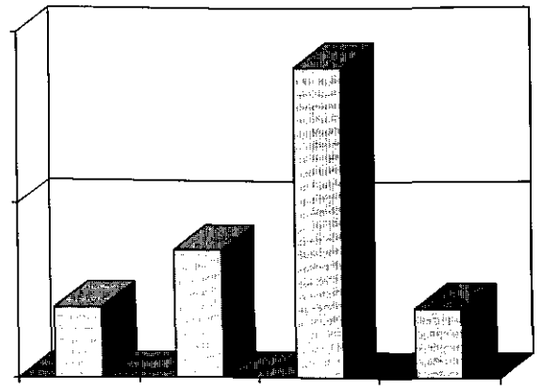


**Figure 19.4** A three-dimensional bar graph plotting the same data as in Figure 19.1.

hard to read, while on the right-hand panel, some poorly chosen gradient fills make the tops of the bars hard to identify. In the first case, different fill patterns would make the comparison of bars clearer, whereas the gradient fills could be fixed by removing them completely, leaving the bars empty.

The choice of fill patterns, etc., will also be influenced by the printers available – be aware that many laser printers don't do a particularly good job of printing solid black, especially if the toner is running low, or if there is wear on some of the internal parts. The same is true of photocopiers, which use the same technology. Unless you are confident that you'll get uniform colours, try using a densely stippled pattern or densely packed cross-hatching. These patterns will print out evenly, even on worn printers. This advice doesn't apply to computer presentations for talks, when solid fills appear much clearer than cross-hatching, etc.

A fault that has been made more common by the availability of graphics software designed for business presentations is the three-dimensional representation of two-dimensional data. This is particularly noticeable for bar graphs and pie charts, although we will emphasize bar graphs here because even two-dimensional pie charts are not much use. A "three-dimensional bar" graph is shown in Figure 19.4. There are many problems with this graph, the most serious being that it is very difficult to tell what value along the *Y*-axis is displayed by the top of the bars.

Note that this is not a three-dimensional graph – only the bars are three-dimensional. The graph in Figure 19.4 shows only two variables, and just adds a third dimension to the graph, without adding any new information. Note also that we haven't even tried to include error bars on this type of graph – the error bars would start somewhere on the tops of each of the bars below, and it would be very hard to see exactly how much overlap there is between means and errors of our groups. As a rule of thumb, or, more usefully, an absolute rule(!):

do **NOT** use three-dimensional graphs for two-dimensional data!!!

There may, however, be occasions when we want to display data with three variables, and may need three axes. Even then, though, it may still be just as good to plot that information in two dimensions. Consider the example on Figure 19.2; it shows the results of measurements on two different factors, but the results can be displayed as a three-dimensional bar graph (Figure 19.5). There is little doubt that the pattern has become less clear. We've now reached almost the peak of obscuring our information, although there is worse to come, and we should also bear in mind Tufte's nomination for the worst graph ever published (Tufte 1983, p. 118).

If we were concerned about the waste of space or ink, we could reduce the simple bar graph even further – a minimalist might argue that we could

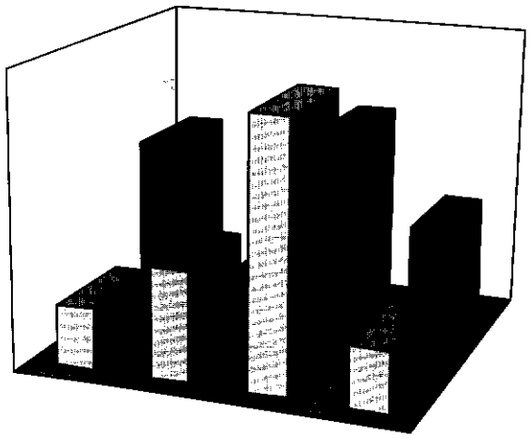


**Figure 19.5** Three-dimensional plot of same data as in Figure 19.2. See how one of the groups is almost completely obscured.

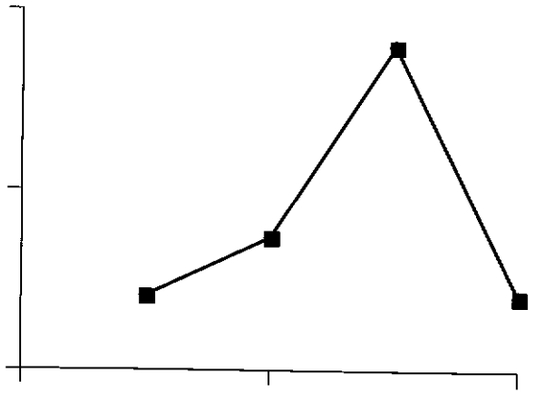


**Figure 19.6** The data from Figure 19.1 displayed as a simple line graph (without error bars, for simplicity).

replace the bar by a single point, without losing any information). In Tufte's terminology, we'd be improving the data:ink ratio – the amount of information conveyed, relative to the amount of ink needed to print it.

### 19.3.2 Line graph (category plot)

Line graphs are like bar graphs except the top of the bar is replaced by a symbol and the adjacent symbols are joined by straight lines (Figure 19.6). They are used when the categorical variable on the *X*-axis can be ordered, or is quantitative, particularly to plot time series. The symbol can represent a single value or the sample mean (or

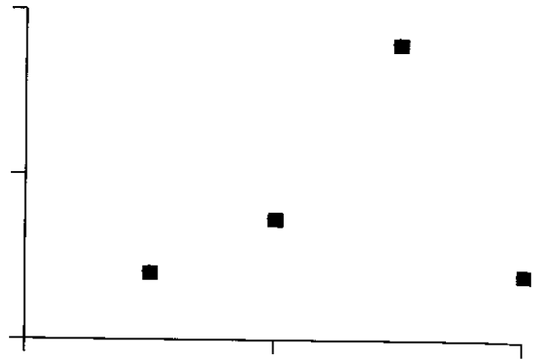


**Figure 19.7** A minimalist graph of the data in Figure 19.6, with the mean of each group now represented by a single point.

median, etc.) – the comments in Section 19.4 about error bars also apply here. These plots are most often used for interaction plots (Chapter 9), and they work very well for this purpose.

It is very important to appreciate that the lines in this case may simply indicate a trend in the (mean) values, without any interpolation. This is particularly the case for interaction plots for fixed effects in analyses of variance – there are by definition no other categories other than those used in the analysis. The line connecting the symbols does not represent any sort of formal relationship between *Y* and *X*, and could be omitted (Figure 19.7).

If we wish to include a second grouping variable, then it can be represented by an additional series of points, with different symbols (or different colors – see Fig. 1 in Cleveland 1994) and/or line styles (Figure 19.8).

### 19.3.3 Scatterplots

We have already discussed scatterplots as an exploratory tool in Chapters 4 and 5. They can also be very effective ways of presenting a bivariate relationship. For example, the scatterplot can include a line that represents a regression or smoothing function fitted to the observations (Figure 19.9). Note that the line in Figure 19.9 extends only to the edge of the range of *X*-values. Many computer graphics packages default to drawing the fitted curve across the entire *X*-axis (see Figure 19.10). This is inappropriate, as we have



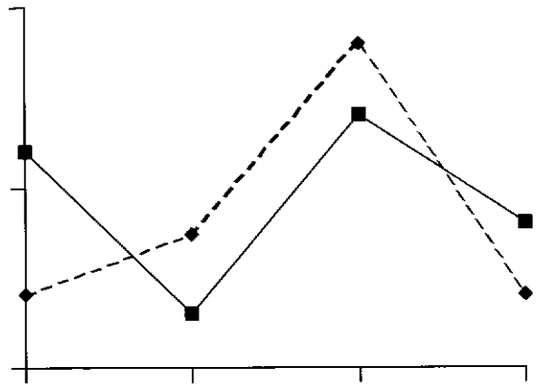


**Figure 19.8** The data from Figure 19.2 plotted as a pair of lines. Compare this with the pseudo-three-dimensional display in Figure 19.5.

no information about the relationship beyond the largest and smallest *X*-values in our sample – even a simple linear relationship might change shape outside our range. A good example of that phenomenon is when we estimate regresssion models for relationships that logically must pass through the origin (e.g. amount of food vs number of limpets $m^{-2}$, mass vs length, etc.), but where the estimated line has a non-zero intercept. The model may be estimated reliably for the range of our data, and because we know that the curve passes through the origin, we therefore know that the line must change slope or shape outside that data range (Figure 19.10; see also Chapter 5).

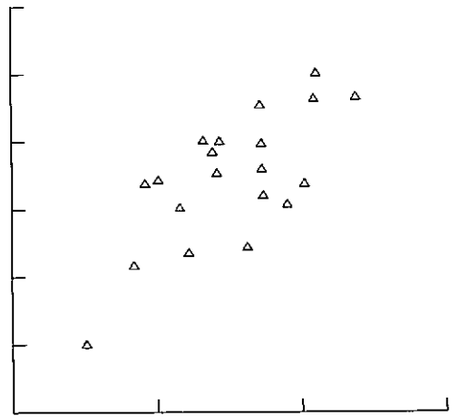


**Figure 19.9** A basic scatterplot, with a least-squares straight line fitted through the observations.

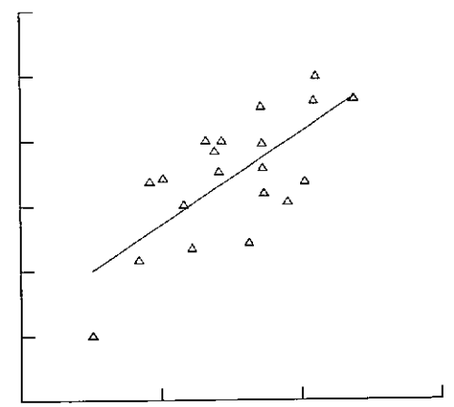
We could also plot confidence intervals about the regression lines or confidence ellipses (Figure 19.11; Sokal & Rohlf 1995) and non-parametric confidence kernels (Silverman 1986) can be included to indicate our level of confidence in the centroid (the mean of the two variables in multidimensional space). Details on these methods were provided in Chapter 5.

Multiple groups can be indicated on the scatterplot by simply using different symbols (or fill patterns or colors) for each group.

### 19.3.4 Pie charts

A pie chart is a circle (or a "pie") where each category's value is represented by a size of its section or slice of the circle (Figure 19.12). The different sections can be further emphasized by different fill patterns or colors.

Pie charts are very commonly used in business graphics (hence their presence in most presentation graphics software) but have a much reduced role in scientific graphics and none in statistical graphics. Tufte (1983) argued that they should never be used because their "data-density" is low and they fail to order numbers along a visual dimension. A reader can't be sure whether to look at the angle or the area to get an idea of how big each group is. Contrast that with a bar or line chart, where there is only one interpretation of the height of the bar or point. It becomes even worse if you allow the software to produce a



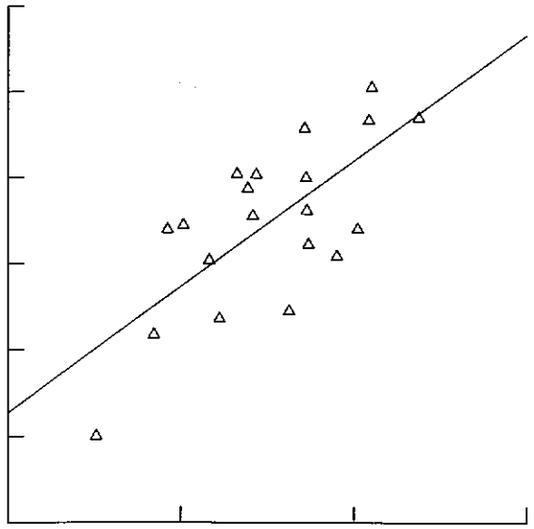


**Figure 19.10** Scatterplot with inappropriate line fitted through the observations.

three-dimensional aspect to this information (Figure 19.13)

## 19.4 Error bars

Any graphical or tabular representation of means should include some measure of the error associated with the estimate of the mean. Common measures of error include the standard deviation (a measure of variability between observations in the sample), the standard error (a measure of precision for the sample mean) and 95% or 99% confidence intervals (Chapter 2). Error bars on graphs are usually represented by a straight line that is symmetrical on either side of the mean. If we are using a bar graph with filled bars, one-sided error bars can be used. The length of the line in each direction indicates one standard deviation or one standard error so the total error bar is two standard errors or, alternatively, the 95% confidence interval.

One problem with error bars on complex graphs with many plotting symbols is that the error bars overlap with each other and other plotting symbols, making the graph messy and difficult to read. In such cases, one alternative is to present the largest and smallest error bars only in one section of the plot to indicate the range of variability or precision in the data.

Where a plot of means relates to a specific analysis, such as a simple ANOVA model, illustrating individual standard deviations or standard errors may not be crucial. In doing the ANOVA, you have assumed that the variances of the different groups are similar and have compared the groups using a pooled estimate of the variation within groups (i.e. the $MS_{Residual}$ term). In showing a single error bar, you may be representing more accurately the variation used in the analysis, whereas the individual errors for the particular treatments may differ from this pooled value, and give the

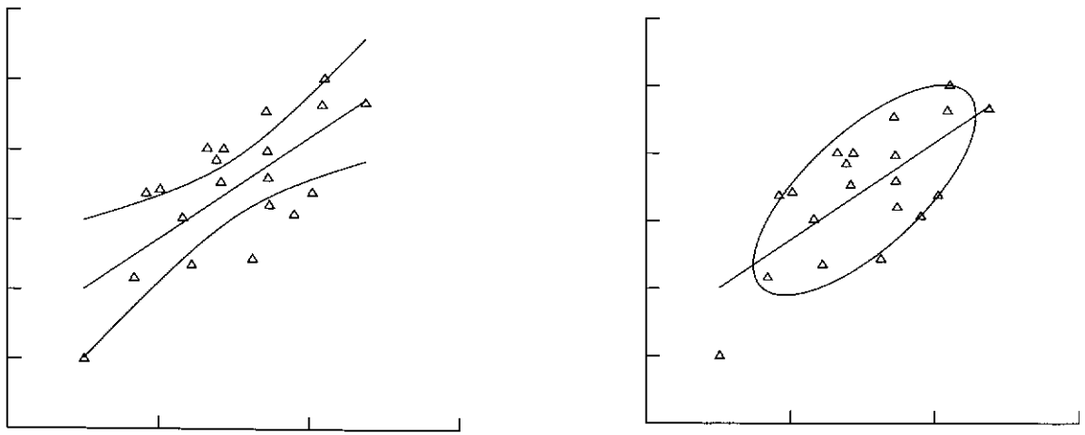


**Figure 19.11** Scatterplots with confidence intervals on the regression line (left) or a confidence ellipse (right).



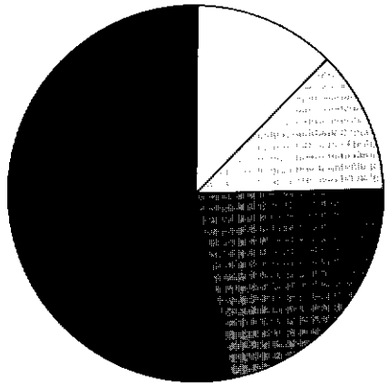


**Figure 19.12** A basic two-dimensional pie chart of the same data as in Figure 19.1.

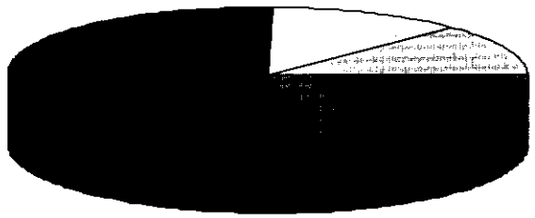


**Figure 19.13** One of the pinnacles of awful graphics, the pseudo-three-dimensional pie chart.

reader an indication of whether the assumption of homogeneous variance is appropriate.

In some more complex linear models, particularly for designs involving nested factors (including repeated measures designs) or combinations of fixed and random factors, a simple standard error or the $MS_{Residual}$ may provide misleading information. As discussed in Chapters 9–12, many different hypotheses are tested in complex models, often using different error terms. As a simple example, consider a two-level nested ANOVA design, with groups as the main factor, plots nested within groups as the nested factor, and replicate observations within plots (Chapter 9). We test the effects of groups against variation among plots within groups, rather than using the within-plots variation. Therefore, if we are describing the differences between groups, we should show some measure of the appropriate variation within groups.

If we use the raw data file (or even the $MS_{Residual}$ from the ANOVA), and plot means and standard errors using common statistical packages, the means may be reliable, but the error bars that are produced by this procedure may bear little relation to the variances used to test particular hypotheses.

The problem is best illustrated with an example. Figure 19.14 shows the graphical summary from three simulated data sets for a nested ANOVA design. All three data sets have four groups, four subgroups within each group, and four replicates per subgroup. The group means were the same across the data sets, as was the variation within subgroups (i.e., the $MS_{Residual}$ was constant). The level of variation between subgroups varied between the data sets, and the graphs show two measures of error.

- The left hand error bar represents the output from a standard statistics package (SYSTAT)[2], from the raw data file. In this figure, the standard error is calculated from all observations within each main group, regardless of the subgroups, i.e., it pools replicate and subgroup variances, and uses the total number of observations in each group as the sample size.
- The right-hand error bar of each pair is based on the variation among subgroups, and was obtained by taking the means for each subgroup, providing a single value for each subgroup, and then plotting means and errors from those data.

The most important thing to note is that the two error bars are similar in some cases, but very different in others, depending on the patterns of variance in a particular data set. When the variation among plots is highest, the "standard" error bars are completely misleading. In the three data sets, the means based on pooling across subgroups will be the same as those calculated from the subgroup means, as long as the number of replicates per subgroup is constant. If the design is unbalanced, the means obtained by the two methods will also be different.

2 The error bars calculated by SYSTAT ignore the structure of the data, and pool all the subgroups into one set of replicates.

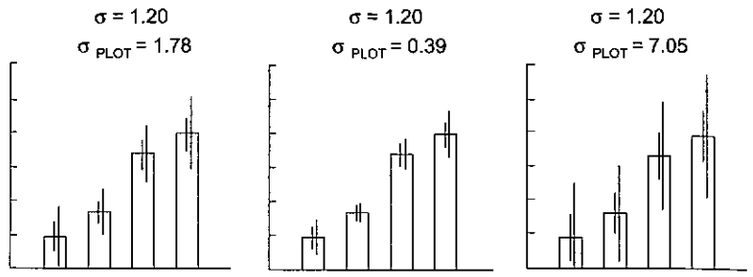



**Figure 19.14** Depiction of error bars for three simulated data sets, from a nested ANOVA design.

This situation becomes more complex if we consider, for example, groups by trials or repeated measures designs (Chapter 11). The test of the an interaction involving the between-subjects and within-subjects factors is made using the variation among subjects across the repeated factor. For example, if the repeated factor is time, the groups by time effect, i.e. the variation in temporal profiles between groups, is tested using the variation through time of the subjects within groups. An overall residual error term will be worthless in this case, and the default output from most software packages would be for error bars to depict the variation among subjects at each level of the within-subjects factor. These error bars might be appropriate for a completely randomized design, but will not have any clear relationship to the denominators used to test the terms of most interest.

### 19.4.1 Alternative approaches

Our strongest recommendation is that you think about the message you want the reader to get, and then think about the measure of variance that is appropriate for this message. The best indication comes from the error term used to test the hypothesis in question, in the case of ANOVA models.

The correct alternative will not always be obvious. To return to the example of the nested ANOVA design, we can identify at least five different error terms that we could calculate.

1. The $\sqrt{MS_{Residual}}$ from the ANOVA.
2. Standard deviations from individual groups, from the raw data file.
3. Standard deviations from a file of means for each subgroup or plot.
4. The $\sqrt{MS_{Subgroup}}$ term from the ANOVA, which averages the variation among subgroups across the groups.
5. The square root of the variance component associated with groups, extracted from the $MS_{Subgroups}$.[3]

As argued earlier, option 1 is incorrect, as is option 2, since they use error terms unrelated to the hypothesis in question. Option 3 provides one correct answer, and results in different error bars for each group. Option 4 is a reasonable approximation, but it leads to an error term that, like option 2, includes two kinds of variation (see Footnote 3). Depending on the relative sizes of the two variances involved, this option may or may not produce an error close to the correct one. Option 5, like option 3, generates an appropriate error, and will produce similar results – given equal sample sizes, it is a pooled estimate of the variation among subgroups, and will be close to the average of the set of subgroup variances.

To see how these options produce different answers, we have used the artificial data sets seen already on Figure 19.14 to produce the data in Table 19.3. You should note that in going from standard deviations to standard errors, options 1 and 2 use the total sample size, i.e. number of subgroups × number of replicates per subgroup. In the above example, with four and four, respectively, the standard errors become quite different. If we take options 3 and 5 as being appropriate, you can see that the other options provide erratic, and misleading error calculations.

Given that most readers tend to look at graphs, and interpret your results for themselves, based on the differences among groups and the error

3 Recall that the Expected Mean Squares for subgroups in a nested ANOVA model is given by $\sigma_e^2 + n\sigma^2_{subgroups}$, and then $\sigma^2_{subgroups}$ can be calculated by $(MS_{Subgroups} - MS_{Residual})/n$.



**Table 19.3** Error bars produced by five different methods, for a two factor nested design, with three different data sets

| Option | Data set 1 | Data set 2 | Data set 3 |
|---|---|---|---|
| 1 | 1.10 | 1.10 | 1.10 |
| 2 | 1.60 | 1.22 | 2.60 |
| 3 | 1.42 | 0.79 | 2.70 |
| 4 | 2.88 | 1.66 | 5.42 |
| 5 | 1.33 | 0.63 | 1.33 |

*Note:*
The numbers are standard deviations.

bars, you risk distracting the readers, or having a reader sceptical of your results, just because you provided other than the most relevant data.

## 19.5 Oral presentations

Although publishing our work in peer-reviewed outlets such as scientific journals is the primary way of making contributions to the field (and of assessing our productivity), talking about our work is a crucial part of publicizing that work, telling colleagues about work in progress, and "advertising" yourself when in the market for scientific jobs. Presenting information clearly and without distractions is just as important for oral papers, with a few additional considerations. There is a range of books and papers offering thoughts on how to construct an effective talk, and, here, we focus on how you display your data and analyses.

Most scientists now prepare talks using a range of graphics packages, most of us lack any training in graphical design, and a substantial number of us have poor taste. These three factors can combine to produce a wide range of distracting graphical displays. While we don't pretend to be style gurus (or may pretend, but unconvincingly!), we can offer some thoughts about preparing audiovisual aids.

### 19.5.1 Slides, computers, or overheads?

One of the first decisions to make is the kinds of tools you'll use to display the information. You will have three main options, assuming that most of us won't use the blackboard for a conference talk or seminar. Computer-based presentations are becoming easier, as more and more venues offer computer projections. Slides remain a very reliable, compact way of presenting information, and offer very high resolution, while overhead projection sheets are completely reliable, and also high resolution.

When deciding which of these you should use, you should consider the following.

- The venue.
  - How big is the room? Many overhead projectors don't produce large images, because they can't be moved far enough from the stage, so you might want to avoid them in big venues.
  - Is the room likely to have good lighting controls? If it can't be darkened, as can be the case at some convention facilities, you may find that your slides can't be seen, and that overheads are much brighter.
  - What is your target audience familiar with? In the past, most people giving talks at scientific conferences used slides. Meetings involving government or industry people, and less formal academic meetings, typically involved overhead projections, with slides being rare. This difference would not affect your preparation of the talk, it was a guide to the kinds of facilities you could expect when you arrived to speak. Now, it is very common to have computer projection facilities available, regardless of the venue.
  - Do you have confidence in the computer facilities? When you turn up for your talk, you may find a beautifully equipped room, with a computer with the latest version of your graphics package, and you'll just need to insert your disk (or even drag your presentation over the internet). Alternatively, you may find that
    - "computer projection facilities" means a plug in the wall, and you are expected to bring a computer with you, or
    - there's an antique Macintosh, when you prepared your talk using the newest

version of a Windows graphics package,or
  - you love an obscure graphics package, used it to prepare your talk, but the computers in the venue lack that program, or
  - you scanned some beautiful images into your presentation, then needed a special high-capacity disk to store your talk. The computer in the room won't read those disks, or
  - ... you can add a range of other, real-world disasters to this list.
- What kind of talk is it? Will you give it once, or is it to be a travelling show that you expect to give a few times, such as a talk about your PhD research? You are likely to give a PhD talk for at least a year or two, and it's probably worth making slides, but, later in your career, you may be asked to give more general or synthetic talks, on a range of topics, and you might write a new talk each time, with little intention of repeating it. One advantage of computer-based presentations is that they can be changed at no cost, while changing a slide costs money and time. It may also be that working in this way encourages you to create a fresh talk, rather than planning your talk based on the slides that you happen to have available.
- How organized are you? If you do everything at the last moment, computer-based presentations offer the most flexibility. It's even possible to change your overheads in response to some profound (or inflammatory) thought offered by the speaker preceding you in the program. You can also fix the spelling error you discovered when running through your talk.
- Where are you going to speak? Slides are the most secure option – they are compact, can be carried with you on planes, aren't affected by magnetic fields, etc. Computer disks are the least stable option, but you can improve things by making sure that you can get another copy of your presentation over the internet if the worst happens, and you can take multiple copies, spread through your baggage, in the heel of your shoe, etc.

### 19.5.2 Graphics packages

Whatever medium you choose, you will almost certainly use one of the common graphics packages to construct your audiovisuals. These packages are written for business users, and the software developers apparently think that business users love to use extraneous, garishly colored graphics as backgrounds. These packages also often lack many of the things we need for scientific purposes – for example, most lack the capacity to plot error bars easily.

We offer a few pieces of opinion (based on extensive, highly selective sampling of our colleagues' biases) about ways to put together a presentation.

- Keep the backgrounds simple. Use a uniform or lightly graded background. Complex, multicolored backgrounds will obscure parts of the text.
- Keep the number of fonts to a minimum.
- Strange transitions between slides – blinds, curtains to the left, checkerboards – and text flying in from all directions can be done easily from most software packages. It tends to polarize your audience. Mixing different transitions and patterns of appearance of objects should be avoided. Doing this demonstrates to the audience that you know how to use the bells and whistles of the software, but it also tells the audience something else about you ... almost certainly an impression you'd like to avoid.

### 19.5.3 Working with color

As a general rule, graphics packages offer sets of colors that are recommended for producing a particular overall look for your presentation.

- We suggest that you choose a particular set, and use exactly those colors, rather than designing your own combination. The color combinations that you select are, with all due respect, likely to be awful, and consist of colors that shouldn't be combined, no matter what your drunken friends think. In addition, many packages offer an option to switch from a color scheme for 35 mm slides or computer graphics (often a dark background and light text), to one designed for overhead projection (a light background, dark text). This switch can be made with a single mouse click, but if you have redefined the color palette, you may lose this ability.
- There is a substantial literature on color perception, and a good understanding of working color combinations. You may want to read some of that literature, and, again, consult Tufte (1983) as an entry point.
- Remember color blindness and its incidence among the general population. There are some color combinations that are offered by many of the graphics packages, particularly red and green, which will be indistinguishable to as many of 20% of your audience (especially if you are in a particularly male-dominated forum).
- As a general rule, use solid fill patterns, and distinguish groups by different colors for audiovisuals (cf. contrasting fill patterns for printed material).

### 19.5.4 Scanned images

To avoid switching between slides and computer, you may decide to scan some images into your presentation. Scanned images are very large, especially if they are stored with fine color detail (e.g. 16.7 million colors on the palette). Individual images can be quite a few megabytes, but you may not need high resolution everywhere.

- If you are converting your presentation into slides, you should scan any images at the highest resolution possible, because 35 mm film is capable of fine details.
- If you are preparing overheads, the resolution will depend on the capabilities of your printer. Use high resolution.
- If you are using computer projection, most systems operate at only $800 \times 600$ pixel resolution. Therefore, if your scanned image exceeds this size, the finer details can't be displayed. You should reduce the resolution to something only slightly finer (i.e. slightly more pixels) than will be displayed. Most images are also scanned with many colors. Reducing the number of colors can dramatically decrease the file size; try reducing the number of colors, and see if the image is degraded. The net result will be a presentation file that is more compact, and fits onto fewer computer disks (at least using twentieth-century technology).

Finally, remember that graphics file types vary in whether they compress the information. Some store the raw graphics information, with no compression. Others compress the file size by searching the image for blocks of identical color, and replacing information about individual pixels with a description of the boundaries of the block and the color. Other file formats, such as JPEG, sacrifice some information for compaction.

### 19.5.5 Information content

Bear in mind that, in a printed paper, we can place large amounts of information on a figure, with the reader having time to digest that information. When presenting the material orally, there's usually less time for the audience to assimilate the information. More importantly, you are speaking more or less continuously, and if you produce an audiovisual with large amounts of information, you'll notice a large part of the audience immediately shift their focus away from you, to concentrate on reading. At that time, you've lost control of the audience, and they won't be listening to you. They may also not be getting the information that you want them to.

In general, you should remove all extraneous information from the figures. As part of your talk, you should guide the audience through the particular figure – show them the key patterns, explain what the different symbols represent, and so on. That way, you control the emphasis that is placed on the information, and the audience feels that they are getting a scientist's view of some information, rather than reading another paper.

You probably do not need to show results of statistical tests on the figure. For example, a regression equation, together with $F$-ratios and $P$-values, adds unneccesary clutter to a scatterplot, and there is often a collective groan in the audience when the next slide is an analysis of variance table. Our strong view is that, ethically, if you talk about a pattern in your data – a difference in groups, a correlation, etc. – you are describing the results of a significant analysis. The audience takes this on trust, and adding the analytical results to your figure or table doesn't help. During the talk, there is no chance to scrutinize your experimental design and analysis, to check that you did everything appropriately, so they must take the analysis on trust, anyway.

## 19.6 General issues and hints

- Presenting results clearly is a neglected part of publicizing scientific work.
- Most statistical packages produce considerable redundancy in their output, and omitting elements of redundancy produces cleaner, more concise, descriptions of results.
- Most graphics packages produce styles of graphs and allow choices of fill pattern and ornamentation that obscure, rather than clarify, results.
- Graphical illustration of results should be tailored to the audience, and optimal use of colors, fill patterns, and explanatory text will be very different for published scientific papers and oral presentations.
- In preparing illustrations, decide what pattern in the data you wish to illustrate, then identify the kind of variation that was the background against which the particular patterns were assessed. This variation is an appropriate candidate for error bars.

# References


Abrahams, M.V. & Townsend, L.D. (1993) Bioluminescence in dinoflagellates: a test of the burglar alarm hypothesis. *Ecology* **74**: 258–260.

Abrams, M.D., Kubiske, M.E. & Mostoller, S.A. (1994) Relating wet and dry year ecophysiology to leaf structure in contrasting temperate tree species. *Ecology* **75**: 123–133.

Agresti, A. (1990) *Categorical Data Analysis*. Wiley, New York.

Agresti, A. (1996) *An Introduction to Categorical Data Analysis*. Wiley, New York.

Aguiar, M.R. & Sala, O.E. (1997) Seed distribution constrains the dynamics of the Patagonian steppe. *Ecology* **78**: 93–100.

Aiken, L.S. & West, S.G. (1991) *Multiple Regression: Testing and Interpreting Interactions*. Sage, Newbury Park.

Akaike, H. (1978) A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics* **30**: 9–14.

Akritas, M.G. (1991) Limitations of the rank transform procedure: a study of repeated measures designs, part 1. *Journal of the American Statistical Association* **86**: 457–460.

Akritas, M.G., Ruscitti, T.F. & Patil, G.P. (1994) Statistical analysis of censored environmental data. In: *Handbook of Statistics Vol. 12 Environmental Statistics* (Patil, G.P. & Rao, C.R. eds.), pp. 221–242. North Holland, Amsterdam.

Allchin, D. (1999) Negative results as positive knowledge, and zeroing in on significant problems. *Marine Ecology Progress Series* **191**: 303–305.

Andersen, P.K. & Keiding, N. (1996) Survival analysis. In: *Advances in Biometry* (Armitage, P. & David, H.A. eds.), pp. 177–199. Wiley, New York.

Anderson, J.L. (1998) Embracing uncertainty: the interface of Bayesian statistics and cognitive psychology. *Conservation Ecology* **2**(2): http://www.consecol.org/vol2/iss1/art2

Anderson, M.J. (2001) A new method for non-parametric multivariate analysis of variance. *Australian Ecology* **26**: 32–46.

Anderson-Sprecher, R. (1994) Model comparisons and $R^2$. *The American Statistician* **48**: 113–117.

Andrew, N.L. & Mapstone, B.D. (1987) Sampling and the description of spatial pattern in marine ecology. *Oceanography and Marine Biology Annual Review* **25**: 39–90.

Andrew, N.L. & Underwood, A.J. (1993) Density-dependent foraging in the sea urchin *Centrostephanus rodgersii* on shallow subtidal reefs in New South Wales, Australia. *Marine Ecology Progress Series* **99**: 89–98.

Anscombe, F.J. (1973) Graphs in statistical analysis. *The American Statistician* **27**: 17–21.

Antelman, G. (1997) *Elementary Bayesian Statistics* (Madansky, A. & McCulloch, R. eds.). Edward Elgar, Cheltenham, UK.

Ayres, M.P. & Scriber, J.M. (1994) Local adaptation to regional climates in *Papilo canadensis* (Lepidoptera: Papilionidae). *Ecological Monographs* **64**: 465–482.

Ayres, M.P. & Thomas, D.L. (1990) Alternative formulations of the mixed-model ANOVA applied to quantitative genetics. *Evolution* **44**: 221–226.

Barnett, V. (1999) *Comparative Statistical Inference*, 3rd edition. Wiley, New York.

Beck, M.W. (1995) Size-specific shelter limitation in stone crabs: a test of the demographic bottleneck hypothesis. *Ecology* **76**: 968–980.

Beck, M.W. (1997) Inference and generality in ecology: current problems and an experimental solution. *Oikos* **78**: 265–273.

Becker, B.J. (1994) Combining significance levels. In: *The Handbook of Research Synthesis* (Cooper, H. & Hedges, L.V. eds.), pp. 215–230. Russell Sage Foundation, New York.

Begon, M., Harper, J.L. & Townsend, C.R. (1996) *Ecology: Individuals, Populations and Communities*, 3rd edition. Blackwell Scientific Publications, London.

Belbin, L. & McDonald, C. (1993) Comparing three classification strategies for use in ecology. *Journal of Vegetation Science* **4**: 341–348.

Belbin, L., Faith, D.P. & Milligan, G.W. (1993) A comparison of two approaches to beta-flexible clustering. *Multivariate Behavioral Research* **27**: 417–433.

Bellgrove, A., Clayton, M.N. & Quinn, G.P. (1997) Effects of secondarily treated sewage effluent on intertidal macroalgal recruitment processes. *Marine and Freshwater Research* **48**: 137–146.

Belsley, D.A., Kuh, E. & Welsch, R.E. (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York.

Bence, J.R. (1995) Analysis of short time series: correcting for autocorrelation. *Ecology* **76**: 628–639.

Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* **B 57**: 289–300.

Berger, J.O. & Berry, D.A. (1988) Statistical analysis and the illusion of objectivity. *American Scientist* **76**: 159–165.

Berger, J.O. & Sellke, T. (1987) Testing a point null hypothesis: the irreconcilability of $P$ values and evidence. *Journal of the American Statistical Association* **82**: 112–122.