Other sampling designs take into account heterogeneity in the population from which we are sampling. Stratified sampling is where the population is divided into levels or strata that represent clearly defined groups of units within the population and we sample independently (and randomly) from each of those groups. For example, we may wish to estimate characteristics of a population of stones in a stream (our variable might be species richness of invertebrates). If the stones clearly fall into different habitat types, e.g. riffles, pools and backwaters, then we might take random samples of stones from each habitat (stratum) separately. Stratified sampling is likely to be more representative in this case than a simple random sample because it ensures that the major habitat types are included in the sample. Usually, the number of units sampled from each stratum is proportional to the total number of possible units in each stratum or the total size of each stratum (e.g. area). Estimating population means and variances from stratified sampling requires modification of the formulae provided in Chapter 2 for simple random sampling. If sampling within a stratum is random, the estimate of stratum population mean is as before but the estimate of the overall population mean is:

$$\bar{y}_{str} = \sum_{h=1}^{l} W_h \bar{y}_h \tag{7.1}$$

where there are $h = 1$ to l strata, W_h is the proportion of total units in stratum h (often estimated from the proportion of total area in stratum h) and $\bar{y}_h$ is the sample mean for stratum h (Levy & Lemeshow 1991). If our sample size within each stratum is proportional to the number of possible units within each stratum, Equation (7.1) simplifies to:

$$\bar{y}_{str} = \frac{\sum_{h=1}^{l} \sum_{i=1}^{n_h} y_{hi}}{n} \tag{7.2}$$

where there are $i = 1$ to n_h observations sampled within stratum h, y_{hi} is the ith observation from the hth stratum and n is the total sample size across all strata. The standard error of this mean is:

$$s_{\bar{y}_{str}} = \sqrt{\sum_{h=1}^{l} (W_h)^2 \frac{s_h^2}{n_h}} \tag{7.3}$$

where s_h^2 is the sample variance for stratum h. Approximate confidence intervals can also be determined (Levy & Lemeshow 1991, Thompson 1992). When statistical models are fitted to data from stratified sampling designs, the strata should be included as a predictor variable in the model. The observations from the different strata cannot be simply pooled and considered a single random sample except maybe when we have evidence that the strata are not different in terms of our response variable, e.g. from a preliminary test between strata.

Cluster sampling also uses heterogeneity in the population to modify the basic random sampling design. Imagine we can identify primary sampling units (clusters) in a population, e.g. individual trees. For each primary unit (tree), we then record all secondary units, e.g. branches on each tree. Simple cluster sampling is where we record all secondary units within each primary unit. Two stage cluster sampling is where we take a random sample of secondary units within each primary unit. Three stage cluster sampling is where we take a random sample of tertiary units (e.g. leaves) within each secondary unit (e.g. branches) within each primary unit (e.g. trees). Simple random sampling is usually applied at each stage, although proportional sampling can also be used. These designs are used to estimate variation at a series of hierarchical (or nested) levels, often representing nested spatial scales and nested linear ANOVA models are often fitted to data from two or more stage cluster sampling designs (Section 9.1).

Systematic sampling is where we choose sampling units that are equally spaced, either spatially or temporally. For example, we might choose plots along a transect at 5 m intervals or we might choose weekly sampling dates. Systematic sampling is sometimes used when we wish to describe an environmental gradient and we want to know where changes in the environment occur. For example, we want to measure the gradient in species richness away from a point source of pollution. Simple random sampling away from the source might miss the crucial region where the species richness undergoes rapid change. Sampling at regular intervals is probably a better bet. Various methods exist for estimating means and variances from systematic sampling, although the estimates are biased unless certain conditions are met (Levy & Lemeshow 1991).

The big risk with systematic sampling is that the regular spacing may coincide with an unknown environmental gradient and so any inference to the whole population of possible sampling units would be biased (Manly 2001). This is probably more likely in field biology (e.g. ecology) where environmental gradients can occur at a range of different spatial and temporal scales.

Systematic sampling can have a single random starting point, where the first unit is chosen randomly and then the remainder evenly spaced. Alternatively, a cluster design could be used, where clusters are chosen at random and then systematic selection on secondary sampling units within each cluster is used.

Finally, we should briefly mention adaptive sampling. When a sampling program has a temporal component, which is often the case in biology, especially when sampling ecological phenomena or environmental impacts, then we might modify our sampling design on the basis of estimates of parameters early in the program. For example, we might change our sample size based on preliminary estimates of variance or we might even change to a stratified design if the initial simple random sampling indicates clear strata in the population that were not detected early on. Thompson (1992) provides an introduction to adaptive sampling but a more detailed text is Thompson & Seber (1995).

7.1.2 Size of sample

If we have idea of the level of variability between sampling units in our population, we can use this information to estimate the required sample size to be confident (e.g. 95% confident) that any sample mean will not be different from the true mean by more than a specified amount under repeated sampling. The calculations are simple, assuming we have sampled randomly and the Central Limit Theorem (Chapter 2) holds:

$$n \geq \frac{z^2 \sigma^2}{d^2} \tag{7.4}$$

where z is the value from a standard normal distribution for a given confidence level (z equals 1.96 for 95% confidence so z^2 approximately equals four – Manly 2001), σ^2 is the variance of the population (usually estimated with s^2 from some pilot sample or previous information) and d is the maximum allowable absolute difference between the estimated mean and the true population mean. Note that the estimation of sample sizes depends on the variance estimate from the pilot study matching the variance in the population when we sample.

7.2 Experimental design

While our emphasis is on manipulative experiments, most of the principles we will outline below also apply to non-manipulative contrasts that we might make as part of sampling programs. General principles of experimental design are described in many standard statistical texts, and in great statistical detail in some very good, specialized books, such as Mead (1988) and Underwood (1997). Hairston (1989) and Resetarits & Fauth (1998) describe many examples of ecological experiments and evaluate their design.

The most important constraint on the unambiguous interpretation of an experiment is the problem of confounding. Confounding means that differences due to experimental treatments, i.e. the contrast specified in your hypothesis, cannot be separated from other factors that might be causing the observed differences. A simple, albeit trivial, example will illustrate the problem. Imagine you wished to test the effect of a particular hormone on some behavioral response of crayfish. You create two groups of crayfish, males and females, and inject the hormone into the male crayfish and leave the females as the control group. Even if other aspects of the design are OK (random sampling, controls, etc.), differences between the means of the two groups cannot be unambiguously attributed to effects of the hormone. The two groups are also different genders and this may also be, at least partly, determining the behavioral responses of the crayfish. In this example, the effects of hormone are confounded with the effects of gender. The obvious solution is to randomize the allocation of crayfish to treatments so that the two groups are just as

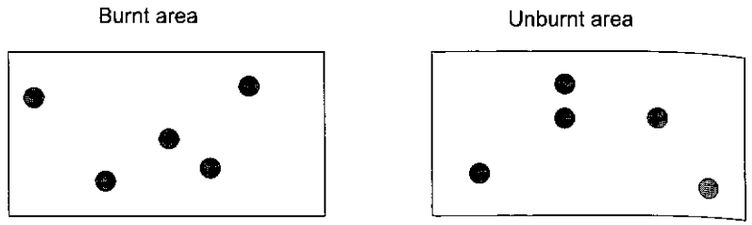

Figure 7.1 Example of an inappropriately replicated study on the effects of fire on soil invertebrates. Each area is sampled with five replicate soil cores.

likely to have males and females. Unfortunately, possible confounding is rarely this obvious and confounding can sneak into an experimental design in many ways, especially through inappropriate replication, lack of proper controls and lack of randomized allocation of experimental units to treatments. These issues will be our focus in this chapter.

Sometimes, confounding is a deliberate part of experimental design. In particular, when we have too many treatment combinations for the number of available replicate units, we might confound some interactions so we can test main effects (Chapter 9). Designs with such deliberate confounding must be used with care, especially in biology where interactive effects are common and difficult to ignore.

7.2.1 Replication

Replication means having replicate observations at a spatial and temporal scale that matches the application of the experimental treatments. Replicates are essential because biological systems are inherently variable and this is particularly so for ecological systems. Linear model analyses of designed experiments usually rely on comparing the variation between treatment groups to the inherent variability between experimental units within each group. An estimate of this latter variability requires replicate units.

Replication at an appropriate scale also helps us avoid confounding treatment differences with other systematic differences between experimental units. For example, to test if there are effects of fish predation on the abundance of a species of bivalve on intertidal mudflats, we might set up a field experiment using fish exclusion cages and suitable cage controls (see Section 7.2.2 for discussion of controls) over plots (experimental units) on the mudflat. If we simply have a single exclusion plot and a single control plot, then the effects of our treatment (fish exclusion) are confounded with inherent differences between the two plots related to their spatial location, such as tidal height, sediment composition, etc. With two or more replicate plots for each of the two treatments (exclusion and control), we can be much more confident in attributing differences between treatment and control plots to fish exclusion rather than inherent plot differences. Note that replication does *not* guarantee protection from confounding because it is still possible that, by chance, all our treatment plots are different from our control plots in some way besides access to fish. However, the risk of confounding is reduced by replication, especially when combined with randomized allocation of treatments to experimental units (Section 7.2.3).

While most biologists are well aware of the need for replication, we often mismatch the scale of those replicates relative to treatments being applied. Probably no other aspect of experimental design causes more problems for biologists (Hurlbert 1984). Imagine a study designed to test the effects of fire on the species richness of soil invertebrates. Fire is difficult to manipulate in the field, so investigators often make use of a natural wildfire. In our example, one burnt area might be located and compared to an unburnt area nearby. Within each area, replicate cores of soil are collected and the species richness of invertebrates determined for each core (Figure 7.1). The mean number of species of invertebrates between the two areas was compared with a *t* test, after verifying that the assumptions of normality and equal variances were met.

There is nothing wrong with the statistical test in this example. If the assumptions are met, a *t* test is appropriate for testing the H_0 that there is no difference in the mean number of invertebrate species between the two areas. The difficulty is that the soil cores are not the appropriate scale of

replication for testing the effects of fire. The spatial unit to which fire was either applied or not applied was the whole area, and the measures of species richness from within the burned area measure the impact of the same fire. Therefore, there is only one replicate for each of the two treatments (burnt and unburnt). With only a single replicate area for each of our treatments, the effect of fire is completely confounded with inherent differences between the two areas that may also affect invertebrates, irrespective of fire. It is very difficult to draw conclusions about the effect of fire from this design; we can only conclude from our analysis that the two areas are different.

The replicate soil cores within each area simply represent subsamples. Subsampling of experimental units does not provide true replication, only pseudoreplication (*sensu* Hurlbert 1984). Pseudoreplication is a piece of jargon that has been adopted by many biologists and used to refer to a wide range of flawed experimental designs. In many cases, biologists using this term do not have a clear understanding of the problem with a particular design, and are using the phrase as a catch-all to describe different kinds of confounding. We will avoid the term, in part to encourage you to learn enough of experimental design to understand problem designs, but also because the term is a little ambiguous. The design is replicated, but the replication is at the wrong scale, with replicates that allow us to assess each area, and the differences between areas, but no replicates at the scale of the experimental manipulation.

Confounding as a result of inappropriate replication is not restricted to non-manipulative field studies. Say as marine biologists, we wished to test the effects of copper on the settlement of larvae of a species of marine invertebrate (e.g. a barnacle). We could set up two large aquaria in a laboratory and in each aquarium, lay out replicate substrata (e.g. Perspex panels) suitable for settling barnacle larvae. We dose the water in one aquarium with a copper solution and the other aquarium with a suitable inert control solution (e.g. seawater). We then add 1000 cyprid larvae to each aquarium and record the number of larvae settling onto each of the panels in each aquarium. The mean number of settled larvae between the two aquaria was compared with a *t* test.

We have the same problem with this experiment as with the fire study. The appropriate experimental units for testing the effects of copper are the aquaria, not individual panels within each aquarium. The effects of copper are completely confounded with other inherent differences between the two aquaria and panels are just subsamples. We emphasize that there is nothing wrong with the *t* test; it is just not testing a null hypothesis about copper effects, only one about differences between two aquaria. To properly test for the effects of copper (rather than just testing for differences between two aquaria), this experiment requires replicate treatment and control aquaria. Note that this experiment has other problems, particularly the lack of independence between the multiple larvae in one aquarium – barnacle cyprids are well known to be gregarious settlers.

As a final example, consider a study to investigate the effects of a sewage discharge on the biomass of phytoplankton in a coastal habitat. Ten randomly chosen water "samples[1]" are taken from the sea at a location next to the outfall and another ten water "samples" are taken from the sea at a location away (upcurrent) from the outfall. As you might have guessed, the appropriate units for testing the effects of sewage are locations, not individual volumes of water. With this design, the effect of sewage on phytoplankton biomass is completely confounded with other inherent differences between the two locations and the water "samples" are just subsamples.

How do we solve these problems? The best solution is to have replicates at the appropriate scale. We need replicate burnt and unburnt areas, replicate aquaria for each treatment, replicate locations along the coast with and without sewage outfalls. Such designs with correct replication provide the greatest protection against

1 Biologists and environmental scientists often use the term sample to describe a single experimental or sampling unit, e.g. a sample of mud from an estuary, a sample of water from a lake. In contrast, a statistical sample is a collection of one or more of these units ("samples") from some defined population. We will only use the term sample to represent a statistical sample, unless there are no obvious alternative words for a biological sample, as in this case.

confounding. In some cases, though, replication is either very difficult or impossible. For example, we might have an experiment in which constant temperature rooms are the experimental units, but because of their cost and availability within a research institution, only two or three are available. In the example looking at the effects of sewage outfalls, we usually only have a single outfall to assess, although there may be no limit to the availability of locations along the coast without outfalls. Experiments at very large spatial scales, such as ecosystem manipulations (Carpenter *et al.* 1995), often cannot have replication because replicate units simply don't exist in nature.

In situations where only one replicate unit is possible for each treatment, especially in a true manipulative experiment that is relatively short-term, one possibility is to run the experiment a number of times, each time switching the treatments between the experimental units. For example, run the copper experiment once, and then repeat it after reversing which aquarium is the treatment and which is the control. Repositioning the aquaria and repeating the experiment a number of times will reduce the likelihood that differences between aquaria will confound the effects of copper. Alternatively, we could try and measure all variables that could possibly influence settlement of barnacles and see if they vary between our aquaria – if not, then we are more confident that the only difference between aquaria is copper. Of course, we can never be sure that we have accounted for all the relevant variables, so this is far from an ideal solution.

For the sewage outfall example, the problem of confounding can be partly solved by taking samples at several places well away from the outfall, so we can at least assess the amount of variation between places. Ideally, however, we need samples from several outfalls and corresponding areas far away, but it is difficult to recommend the installation of multiple outfalls just for statistical convenience. A substantial literature has developed to try and make a conclusion about impacts of human activities when there is only one place at which a potential impact occurs. These designs are generally called Before-After-Control-Impact (BACI) designs (Green 1979, Stewart-Oaten *et al.* 1986), and various suggestions include sampling through time to provide replication, sampling multiple control areas, etc. These designs have been contentious, and a critical evaluation of their pros and cons can be found in Keough & Mapstone (1995) and Downes *et al.* (2002).

The above examples illustrate spatial confounding, but confounding with time can also occur, although it is less common. Consider an experiment to test for the effects of floods on drifting insects in streams. We might set up six artificial stream channels with drift nets at the end – six stream channels are all we have available. We want to impose two treatments, high flow and normal flow, and we know from previous work that we will need a minimum of six replicates per treatment to detect the desired effect if it occurs (see Section 7.3 on power analyses). We could do the experiment at two times with six replicates of high flow at time one and six replicates of normal flow at time two. Unfortunately, the effects of flow would be completely confounded with differences between the two times. The appropriate design of this experiment would be to have three replicates of each treatment at each time, therefore becoming a two factor experiment (treatment and time). If we only have enough experimental units to have one replicate for each treatment, then we can use time as a blocking factor (see Chapter 10).

7.2.2 Controls

In most experimental situations, many factors that could influence the outcome of the experiment are not under our control and are allowed to vary naturally. Therefore, it is essential to know what would happen if the experimental manipulation had not been performed. This is the function of controls. An excellent example of the need for controls comes from Hairston (1980, see also 1989) who wished to test the hypothesis that two species of salamanders (*Plethodon jordani* and *P. glutinosus*) in the Great Smoky Mountains compete. He set up experiments where *P. glutinosus* was removed from plots. The population of *P. jordani* started increasing during the three years following *P. glutinosus* removal, but the population of *P. jordani* on control plots (with *P. glutinosus* not removed) showed an identical increase. Without the control plots, the increase in *P. jordani* might have been incorrectly attributed to *P. glutinosus* removal.

Simply deciding to have controls is not enough. The controls must also allow us to eliminate as many artifacts as possible introduced by our experimental procedure. For example, research in animal physiology often looks at the effects of a substance (e.g. some drug or hormone or toxin) on experimental animals, e.g. rats, or *in vitro* tissue preparations. The effects of the substance are assessed by comparing the response of animals injected with the substance to the response of control animals not injected. However, differences in the responses of the two groups of animals may be due to the injection procedure (handling effects, injury from needle etc.), not just the effect of the substance. The effects of the substance are confounded with differences in experimental procedure. Such an experiment would need control animals that are injected with some inert substance (e.g. saline solution), but which undergo the experimental procedure identically to the treatment animals; such a control is sometimes termed a procedural control. Then any difference between the groups can be more confidently attributed to the effect of the substance alone.

Ecological field experiments also offer challenges in designing appropriate controls (Hairston 1989, Underwood 1997). For example, to examine the effect of predatory fish on marine benthic communities, we might compare areas of substratum with fish exclusion cages to areas of substratum with no cages. However, the differences between two types of area may be due to effects of the cages other than excluding fish (e.g. shading, reduced water movement, presence of hard structure). The effects of fish exclusion are confounded with these other caging effects. We must use cage controls, e.g. cages that have larger gaps in the mesh that allow in fish but are otherwise as similar to the exclusion cages as possible. Then, any difference between treatments can be more confidently attributed to the effect of excluding fish alone. This is not a simple matter – if a major effect of cages is to alter water movement (and hence sedimentation), it may be difficult to leave big enough gaps for fish to enter at the same rate as they enter uncaged areas, without changing flow rates. In many cases, the cage control will be physically intermediate between caged and uncaged areas. The marine ecological literature contains many examples of different kinds of cage controls, including the step of using cages to both enclose and exclude a particular predator.

Ecological experiments sometimes involve translocating organisms to different areas to test a specific hypothesis. For example, to test what determines the lower limit of intertidal gastropods on intertidal rocky shores, we might consider translocating gastropods to lower levels of the shore. If they die, it may be an effect of height on the shore or an effect of translocation procedure. Appropriate controls should include gastropods that are picked up and handled in exactly the same way as translocated animals except they are replaced at the original level. Additional controls could include gastropods at the original level that are not moved, as a test for the effects of handling by themselves. Controls for translocation experiments are tricky – see Chapman (1986) for a detailed evaluation.

7.2.3 Randomization

There are two aspects of randomization that are important in the design and analysis of experiment. The first concerns random sampling from clearly defined populations, as we discussed in Chapter 2 and in Section 7.1.1. It is essential that the experimental units within each of our treatments represent a random (or at least haphazard) sample from an appropriate population of experimental units. This ensures that our estimates of population parameters (means, treatment effects, mean squares) are unbiased and our statistical inferences (conclusions from the statistical test) are reliable.

For example, our experimental animals that received a substance in a treatment should represent a random sample of all possible animals that we could have given the substance and about which we wish to draw conclusions. Our caged plots in the marine example must be a random sample of all possible caged plots in that habitat – similarly for our control plots. We must clearly define our treatment (and control) populations when we design our experiment. The converse is

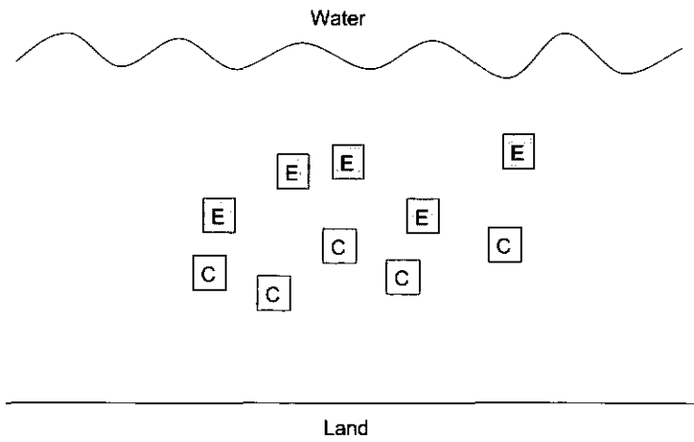

Figure 7.2 Possible result of random allocation of ten plots on an intertidal mudflat to two treatments – fish exclusion (E) and cage-control (C).

that we can only draw conclusions about the population from which we have taken a random sample. If our plots on a mud flat were scattered over a 20 m × 20 m area, then our conclusions only apply to that area; if we used a particular strain of rats, then we have only a conclusion about that genetic strain, and so on.

The second aspect of randomization concerns the allocation of treatments to experimental units or vice versa. One of the standard recommendations in experimental design is that the experimental units be randomly allocated to treatment groups. This means that no pattern of treatments across experimental units is subjectively included or excluded (Mead 1988) and should ensure that systematic differences between experimental units that might confound our interpretation of treatment effects are minimized (Hurlbert 1984, Underwood 1997). The crayfish example described at the beginning of Section 7.2 is an illustration, if somewhat contrived, of the problem.

An artificial example, analogous to one described by Underwood (1997), involves an experiment looking at the difference in growth rates of newly hatched garden snails fed either the flowers or the leaves of a particular type of plant. The flowers are only available for a short period of time, because the plant flowers soon after rain. When the flowers are available, we feed it to any snails that hatch over that period. Snails that hatch after the flowering period are given the leaves of the plant. The obvious problem here is that the two groups of snails may be inherently different because they hatched at different times. Snails that hatch earlier may be genetically different from snails that hatch later, have had different levels of yolk in their eggs, etc. Our results may reflect the effect of diet, or they may reflect differences in the snails that hatch at different times, and these two sources of variation are confounded. Clearly, we should take all the snails that hatch over a given period, say the flowering period, and give some of them flowers and others leaves to eat.

The allocation of experimental units to treatments raises the difficult issue of randomization versus interspersion (Hurlbert 1984). Reconsider the experiment described earlier on the effects of fish predation on marine benthic communities. Say we randomly choose ten plots on an intertidal mudflat and we randomly allocate five of these as fish exclusion (E) plots and five as cage-control (C) plots. What do we do if, by chance, all the control plots end up higher on the shore than all the exclusion plots (Figure 7.2)? Such an arrangement would concern us because we really want our treatment and control plots to be interspersed to avoid confounding fish effects with spatial differences such as tidal height. The simplest solution if we end up with such a clumped pattern after an initial randomization is to re-randomize – any other pattern (except the complete reverse with all control plots lower on the shore) will incorporate some spatial interspersion of treatments and controls. However, we must decide *a priori* what degree of spatial clumping of treatments is unacceptable; re-randomizing until we get a particular pattern of interspersion is not really randomization at all.

Why not guarantee interspersion by arranging

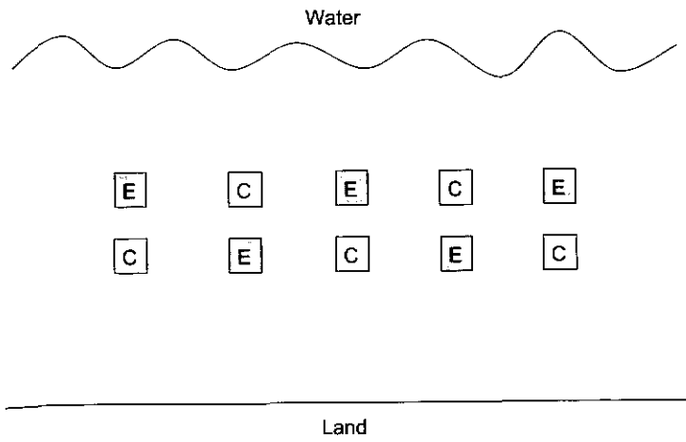

Figure 7.3 Regular positioning of ten plots on mudflat combined with systematic allocation of plots to two treatments – fish exclusion (E) and cage-control (C) – to guarantee interspersion.

our plots regularly spaced along the shore and alternating which is exclusion and which is control (Figure 7.3)? One problem with this design is that our plots within each group no longer represent a random sample of possible plots on this shore so it is difficult to decide what population of plots our inferences refer to. Also, it is possible that the regular spacing coincides with an unknown periodicity in one or more variables that could confound our interpretation of the effects of excluding fish. A compromise might be to randomly select plots on the shore but then ensure interspersion by alternating exclusions and controls. At least we have chosen our plots randomly to start with so the probability of our treatments coinciding with some unknown, but systematic, gradient along the shore won't change compared to a completely randomized design. There is still a problem, however; because, once we have allocated an E, the next plot must be a C, and it becomes more difficult to know what population our E and C plots refer to. This example has additional complications – our replicates will not be truly random, as we will have some minimal separation of replicates. We would not place plots on top of each other, and, as biologists, we have some feeling for the distance that we need to keep plots apart to ensure their independence. If the minimum separation distance is large, we may tend towards uniformly spaced replicates. In a field study, it is also possible that plots are easier to find when they are regular, or, for example if we are working on an intertidal mudflat, with plots not marked clearly, regular spacing of plots makes it easier for researchers and their assistants to avoid walking on one plot accidentally when moving across the area. The eventual positioning of replicates will be a combination of desired randomization, minimum spacing, and logistic considerations.

This issue of randomization versus interspersion illustrates one of the many grey areas in experimental design (and in philosophy – see debate between Urbach 1984 and Papineau 1994). Randomization does not guarantee avoidance of confounding but it certainly makes it less likely. With only a small number of experimental units, spatial clumping is possible and deliberate interspersion, but combined with random sampling, might be necessary. It is crucial that we recognize the potential problems associated with non-randomized designs.

7.2.4 Independence

Lack of independence between experimental units will make interpretation difficult and may invalidate some forms of statistical analysis. Animals and plants in the same experimental arena (cage, aquarium, zoo enclosure, etc.) may be exposed to a set of physical and biological conditions that are different from those experienced by organisms in other arenas. We may have a number of preparations of tissue from a single animal, and other such sets taken from other animals. The animals may differ from each other, so two tissue samples from the same animal might have more similar responses than two pieces of tissue chosen at random from different animals or plants. We will consider statistical problems arising from lack of independence in the appropriate chapters.

7.2.5 Reducing unexplained variance

One of the aims of any biological research project is to explain as much about the natural world as possible. Using linear models, we can estimate the amount of variation in our response variable that we have explained with our predictor variables. Good experimental design will include consideration of how to reduce the unexplained variation ($MS_{Residual}$) as much as possible. There are two broad strategies to achieve this.

- Including additional predictor variables in our analyses. We have discussed this in the context of multiple regression in Chapter 6 and will examine it further in the analysis of multifactor experiments in Chapter 9.
- Change the spatial structure of the design, particularly by incorporating one or more blocking variables. This will be discussed in Chapters 10 and 11.

7.3 | Power analysis

Recall from Chapter 3 that the complement to a Type II error is the concept of power – the long-run probability of detecting a given effect with our sample(s) if it actually occurs in the population(s). If β is the risk of making a Type II error, $1-\beta$, or power, is the probability that we haven't made an error. More usefully, statistical power is a measure of our confidence that we would have detected an important effect if one existed.

This concept can be used in a range of situations. In designing an experiment or making an *a posteriori* assessment of the usefulness of an experiment, the important questions are as follows.

Supposing that there *is* a change of a particular size, what kind of sampling program would be needed to detect that change with reasonable certainty (or to estimate the magnitude of such a change)? Or, given a particular level of resources, what kind of change could we reasonably expect to detect? For *post hoc* assessment (of a non-significant result), we must ask, if our treatments really did have an effect (of a particular size), would we have detected that effect with our experimental design and analysis?

Power analysis is therefore a useful tool for designing an experiment, and it should (but will not, unfortunately, in many cases) also provide justification for publishing non-significant results.

An emerging body of the statistical and biological literature is concerned with questions of power. Here we provide a very broad overview of the uses of statistical power, but for detailed planning of specific experiments or programs, good general reviews are provided by Cohen (1988, 1992), Peterman (1990a,b), National Research Council (1990), Fairweather (1991), and Keough & Mapstone (1995). We will also return to power analysis as we begin to consider more complex designs later in this book.

To determine the power of an analysis, we need to specify the alternative hypothesis (H_A), or effect size, that we wish to detect. For most types of analyses (e.g. simple two group comparisons, ANOVA and regression models), power is proportional to the following.

- Effect size (ES) – how big a change is of interest. We are more likely to detect large effects.
- Sample size (n) – a given effect is easier to detect with a larger sample size.
- Variance (σ^2) between sampling or experimental units – it is harder to detect an effect if the population is more variable.
- Significance level (α) to be used. Power varies with α. As mentioned in Chapter 3, most biologists use a value of $\alpha = 0.05$.

More formally,

$$\text{Power} \propto \frac{ES\ \alpha\ \sqrt{n}}{\sigma} \qquad (7.5)$$

Exactly how we link values of these parameters to power depends on the particular statistical test being used (hence the proportional sign in the equation). For individual cases, we construct a specific equation, usually using the relevant non-central statistical distribution[2], which in turn requires precise knowledge of the statistical test that will be used (see Box 7.1 and Figure 7.4).

2 A non-central distribution describes the distribution of our test statistic that would be expected if H_A, rather than H_0, is correct.

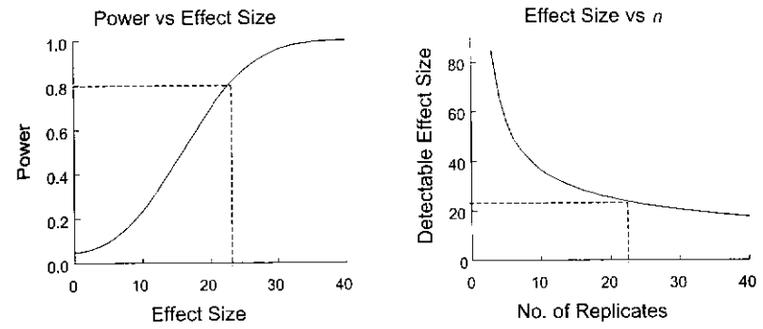

Figure 7.4 Power functions for the abalone example. The panel on the left is for $n = 24$, and the dashed lines indicate the solution for 80% power. The right panel shows detectable effect size vs sample size, and the dashed line shows the calculated effect size.

Box 7.1 | Simple example of power analysis

In an earlier project (Keough & King 1991), we were examining the effect of closing a range of shores to collection of abalone, through proclamation of two marine parks. The closure was contentious, denying both commercial and recreational divers the chance to collect, and, therefore, it was imperative to collect information to test whether the management strategy had worked. The assumption (untested) was that exploitation of abalone had reduced abundances. The intention was to survey a range of rocky headlands after a few years of protection, surveying areas where collection was still allowed and areas where it had been banned (there were no differences between these areas before proclamation of the marine parks). The important question was the feasibility of these surveys. The parameters of the power equation were estimated as follows:

- the test of management could be simplified to a t test, with a replicate observation being a rocky reef site (with some replicate observations within each reef, to get a better idea of its state),
- α was left at 0.05, and $1-\beta$ set to 0.80, by convention, and
- σ was estimated by sending teams of divers out to sample a range of sites in the same way planned for the real monitoring. Those pilot surveys produced a mean density of abalone of 47.5 legal-sized animals per 50 m^2 area, with a standard deviation of 27.7. This latter value was used as an estimate of σ.

In the first case, let's calculate the number of observations (sites) required. Determining the effect size was very difficult, as little work had been done on these animals in the areas concerned, and was eventually calculated using a range of unconnected data sets. As a working assumption, recreational divers and poachers were assumed to take approximately as many animals as commercial divers. Commercial divers were required to file regular reports listing the mass of abalone taken, broken down into small reporting regions. An earlier paper (McShane & Smith 1990) had described size–frequency relationships for commercial catches of abalone, and length–weight relationships (McShane *et al.* 1988), so it was possible to convert a mass of abalone into an average number of animals taken per year from each reporting region. Another fisheries publication provided maps of major abalone reefs, giving their approximate areas. From these data, the number of animals taken could be converted into an approximate number per 50 m^2. In this case, the value for heavily fished areas (averaged over 6 years of diver returns) was

11.6 animals m^{-2}, or approximately 25% of the standing stock. Adding supposed recreational and poaching catches, these values become 23.2, and 50%, respectively.

The power calculations then become quite simple, and can be done using a range of software packages. For these values, the number of sites to be sampled is 24.

In the second case, if we are unhappy with the number of approximations made in calculating the effect size, we could construct a curve of MDES vs n (Figure 7.4)[3]. The relationship is also shown as a curve of power vs effect size for n equals 24, to illustrate the comparison with the first approach. Note that the solution for 80% power corresponds to an effect size of 23.

The important panel is the one for detectable effect size vs sample size, showing that small numbers of sites (less than seven) would require at least a doubling of the number of legal-sized abalone in the area for an effect to show up, whereas our best guess is that the change is more likely to be around 50%, and the dashed line shows that an effect size of 23 corresponds to $n = 24$. The curve also emphasizes the rapid returns resulting from an increase in sample size, if you start with a poorly replicated experiment – the detectable effect declines dramatically at low n values, but tapers off, indicating a region of diminishing return.

3 We constructed the curve shown on using the free software package Power Pack, written by Russell Lenth. His web site (www.divms.uiowa.edu/~rlenth/Power) includes several options for doing power calculations.

7.3.1 Using power to plan experiments (*a priori* power analysis)

There are two ways that power analysis can be used in the design of an experiment or sampling program.

Sample size calculation (power, σ, α, ES known)

The most common use of power analysis during the planning of an experiment is to decide how much replication is necessary. We can then decide whether it is feasible to use this many replicates. To do these calculations, we need to specify the effect size and have an estimate of σ. At the planning stage, you may not have a good idea of the variation you are likely to get, and need to get an estimate, either from previous studies or pilot work. The most difficult step will be specifying the effect size (Section 7.3.3).

Effect size (power, n, σ, known)

If external factors are likely to restrict the number of observations (sample size) to relatively low levels, the alternative approach is to calculate the constraints of the experiment – using this many observations, and with the likely background variability, what is the smallest change that we could expect confidently to identify? This situation is common when the sampling itself is expensive. For example:

- expensive laboratory analyses for trace chemicals,
- benthic marine sampling requiring large ships,
- if there are few laboratories capable of doing assays,
- if processing each observation takes a large amount of your time,
- experimental units are expensive, such as doing physiological work on small mammals, where the cost of each animal may be very restrictive, especially for students.

At either the planning stage, or after an experiment or sampling program has been completed, it is possible to calculate the size of change that could be or could have been detected. This has been termed "reverse power analysis" by Cohen (1988), and the effect size that we calculate has been labelled the Minimum Detectable Effect Size (MDES). We are asking, for a given level of

background variability, sample size, and a desired certainty or power, how big would the change need to be before we would detect it as significant? Again, it is best to use this calculation beforehand, to decide if the work is worth doing, and although it might also be used afterwards to reassure readers that everything was done properly. Calculating the detectable effect may be a preferred solution when you are not comfortable with specifying an *a priori* effect size.

For example, from surveys of intertidal molluscs in protected and collected areas near Williamstown in Victoria, we found changes of 15–25% in the mean size of species that are collected by humans (Keough *et al.* 1993). Because these data came from surveys, rather than controlled experiments, we also measured sizes of a set of species that are not collected by humans in great numbers. To be confident that the patterns seen for collected species did not reflect a response to some unmeasured environmental variable, we analysed the non-collected species, and found no significant difference between sites with and without human access. For non-collected species to be an appropriate control, we need to be confident that we could have detected a pattern the same as that shown by collected species. We used power analysis to show that our sampling program would have detected a change as small as 10% for some of these species, i.e., if non-collected species changed as much as collected ones, we would have detected it (Keough *et al.* 1993).

Sequence for using power analysis to design experiments

The statistical design stage of any experiment or sampling program should include the following steps.

1. State clearly the patterns to be expected if no effect occurs, and the patterns expected if there are changes. In formal statistical terms, this corresponds to clear formulations of the null hypothesis and its alternative.

2. Identify the statistical model to be applied to the data, and state the desired power and the significance levels to be used.

3. Identify the assumptions of that statistical procedure. If possible, use existing or comparable data as a rough guide to whether those assumptions are likely to be satisfied. Consider possible data transformations. If you expect to use transformed data, the effect size must be expressed on that transformed scale. For example, if you are interested in a doubling of numbers of a particular organism, and will analyze log-transformed data, your effect size will be 0.301 when converted to a $\log_{10}$ scale.

4. Obtain some pilot estimate of variation in the variable to be analyzed. In some cases, we require estimates of variation in space and time, while in other cases we may only be comparing in space or through time alone. In some ecological studies, estimating variation through time requires pre-existing data sets involving time periods of at least a few years. If there are no local data, some ballpark estimates may be obtained from the literature from other geographic regions. It is crucial that the estimate of variability must be based on same scales of space and time as your final data. There is no reason to expect that variation on one scale will be a good predictor of variation on a different scale.

If you have complex experimental designs (e.g. Chapters 9–11), you need to think about the variation that is used to test a particular hypothesis. If you have, for example, nested or split-plot designs, different hypotheses will be tested against different measures of variation, and you would need to do power analyses for each separate hypothesis. Importantly in this context, you must get an estimate of σ at the appropriate level.

5. The next step depends on whether your design will be limited by logistical constraints.

(a) If our aim is to design the best possible experiment, we should specify the effect size that we wish to detect – how large a change is of biological interest? The implication here is that detecting changes less than the specified amount has low priority. In practice, this decision is very difficult, but it is nevertheless critical. The effect size may be chosen from a range of sources, e.g. other studies of the same biological system, studies of other processes that you might wish to compare to the one you are investigating, etc. (Section 7.3.3). Using our

desired ES, an estimate of σ and the specified value of α, it should then be possible to calculate the number of replicates needed to detect that effect size with power $1-\beta$.

(b) If we have constraints on the size of our experiment or sampling program, we can use an estimate of σ, the chosen values of α and β and the upper limit to the number of observations possible to determine the Minimum Detectable Effect Size (MDES). It is often useful to calculate MDES values for a range of sampling efforts, and to represent the results as a plot of MDES versus sample size. This relationship can then be used to show how much return we would get for a big change in sampling effort, or the sample size necessary to reach a particular MDES value (see Peterman 1989).

7.3.2 *Post hoc* power calculation

If an experiment or sampling program has been completed, and a non-significant result has been obtained, *post hoc* power analysis can be used to calculate power to detect a specified effect, or to calculate the minimum detectable effect size for a given power. Calculating *post hoc* power requires that we define the effect size we wished to detect, given that we know n and have an estimate of σ. Obviously, once the experiment has been done, we have estimates of σ, e.g. from the $MS_{Residual}$ from a regression or ANOVA model, and we know how much replication we used. The effect size should be the size of change or effect that it is important for us to detect. It is obviously useful to demonstrate that our test had high power to detect a biologically important and pre-specified effect size (Thomas 1997). The downside is that if power is low, all that you have demonstrated is your inability to design a very good experiment, or, more charitably, your bad luck in having more variable data than expected! It is far more useful to use these calculations at the planning stage (Section 7.3.1; Underwood 1999). After an experiment, we would expect to use the calculations to satisfy ourselves that power is high enough, that our initial power calculations, often based on very rough estimates of variance, were correct.

Some statistical packages offer a flawed kind of *post hoc* power calculation, sometimes called "observed power" (Hoenig & Heisey 2001). In this approach, we use the existing analysis to estimate both the effect size and sample variance, and use those values in the power equation. For example, in a two-sample t test, we would use the difference between the two means as the effect size. This observed effect size is unlikely to match a difference that we decide independently is important. Perhaps most importantly, Hoenig & Heisey (2001) have demonstrated that observed power has a 1:1 relationship with the P value so higher P values mean lower power and calculation of observed power tells us nothing new (see also Thomas 1997). We emphasize again the importance of thinking carefully about the kinds of effects that you wish to detect in any experiment, and the value of making this and other decisions before you sample.

Post hoc power calculations can be used to convince reviewers and editors that our non-significant results are worth publishing. Despite the clear value of a confident retention of a null hypothesis (see Underwood 1990, 1999), it can still be difficult in practice to get such results published. We have already emphasized in Chapter 3 that any assessment of the literature can be seriously compromised by the "file-drawer problem". If non-significant results are less likely to be published, because of an active policy of editors and referees or lack of enthusiasm of the researchers, then unbiased syntheses of a particular discipline are not possible. Providing measures of observed effect size and showing you had good power to detect pre-specified effect sizes of biological interest will make non-significant results much more interpretable.

7.3.3 The effect size

The most difficult step of power analyses is deciding an effect size. Our aim is to identify an effect of experimental treatments that we consider important, and that, therefore, we would want to detect. How do we decide on an important effect? The decision is not statistical, but in most cases uses biological judgment by the research worker, who must understand the broad context of the study. In most pieces of research, the work is not self-contained, but our aim is to investigate a phenomenon and to compare that phenomenon to related ones. We might want to:

- compare results for our species to those for other species,
- compare the role of a particular biological process to other processes acting on a particular species or population, or
- contrast the physiological responses to a chemical, gas mixture, exercise regime, etc., to other such environmental changes.

In these cases, we should be guided by two questions. Can we identify a change in the response variable that is important for the organism, such as a change in a respiration parameter, blood pressure, etc., that would be likely to impair an organism's function, or a change in population density that would change the risk of local extinction? What were the levels of response observed in the related studies that we intend to compare to our own? These questions sound simple, but are in practice very difficult, especially in whole-organism biology, where we are often dealing with biological systems that are very poorly studied. In this case, we may not be able to predict critical levels of population depletion, changes in reproductive performance, etc., and will have very little information with which to make a decision. The available information gets richer as we move to sub-organismal measurements, where work is often done on broadly distributed species, standard laboratory organisms, or on systems that are relatively consistent across a wide range of animals or plants. In any case, we must decide what kind of change is important to us.

What if we can not identify an effect size about which we feel confident?

Quite often, we will not be able to select an effect size that we could defend easily. In this case, there are three options available.

1. Use an arbitrary value as a negotiating point. In many published ecological studies, including a range of environmental impact studies, an arbitrary change, usually of 50 or 100% (relative to a control group) in the abundance of a target species, has been used. These values seem to be accepted as being "large", and with the potential to be important. They are not necessarily biologically meaningful – a much smaller change may be important for some populations, while others that vary widely through time may routinely change by 50% or more between years or places. The major value of this approach is in environmental monitoring, where a sampling program may be the result of negotiation or arbitration between interested parties arguing for increases and decreases in the scope of the monitoring program.

2. Cohen (1988) proposed conventions of large, medium, and small effects. Rather than expressing an effect size as, for example, a difference between two means, he standardized the effect size by dividing by σ. For a simple case of comparing two groups, he suggested, based on a survey of the behavioral and psychological literature, values of 0.2, 0.5, and 0.8 for standardized differences (i.e., $(\bar{y}_a - \bar{y}_b)/\sigma$, for small, medium, and large). He acknowledged that these values are arbitrary, but argued that we use arbitrary conventions very often, and proposed this system as one for dealing with cases where there is no strong reason for a particular effect size. These values may or may not be appropriate for his field of research, but they are not necessarily appropriate for the range of biological situations that we deal with. A critical change in migration rates between geographically separated populations, for example, will be very different when we are investigating genetic differentiation between populations, compared to measuring ecologically important dispersal that produces metapopulations. Considerable exchange is necessary for ecological links, but very low rates of exchange are sufficient to prevent genetic separation. Any broad recommendation such as Cohen's must be tempered by sensible biological judgment.

3. A more useful approach may be the one we describe above, in which, rather than use a single effect size, we plot detectable effect size versus sampling effort or power versus effect size. In this case, we get an idea of the kinds of changes that we could detect with a given sampling regime, or, the confidence that we would have in detecting a range of effects. While we don't have a formal criterion for deciding whether to proceed, this approach is useful for giving an idea of the potential of the experiment.

Environmental monitoring – a special case

One common activity for biologists is to assess the effects of various human interventions in the natural environment, and, in this case, we are not always comparing our results to a broader literature, but collecting information to make decisions about the acceptability of a particular activity, in a particular region. The question, then, is whether the activity in question has an unacceptable impact. We need to decide how big a change in the response variable is unacceptable. In this case, we may get advice on the effect size from formal regulations (e.g. criteria for water quality, setting standards for human health or environmental "health"). There may also be occasions when the level at which the human population becomes concerned defines the target effect size. This level may be unrelated to biological criteria. For example, oiled seabirds washing up on beaches triggers public complaints, but the number of sick or dead animals may not result in a population decline. There will, however, be intense pressure to monitor charismatic megafauna, with an effect size determined by political considerations. In other monitoring situations, we may fall back on arbitrary values, using them as a negotiating point, as described above. Keough & Mapstone (1995, 1997) have described this process, and there is a good discussion of effect sizes in Osenberg *et al.* (1996).

7.3.4 Using power analyses

The importance of these power calculations is that the proposed experiment or sampling program can then be assessed, to decide whether the MDES, power, or sample size values are acceptable. For example, if the variable of interest is the areal extent of seagrass beds, and a given sampling program would detect only a thousand-fold reduction over ten years, it would be of little value. Such a reduction would be blindingly obvious without an expensive monitoring program, and public pressure would stimulate action before that time anyway.

If the results of the power analyses are acceptable because the MDES is small enough, or the recommended number of observations is within the budget of the study, we should proceed. If the solution is unacceptable, the experiment will not be effective, and the level of replication should be increased. If you decide to go ahead with no increase in sample size, it is important that you are aware of the real limitations of the sampling. Proceeding with such a program amounts to a major gamble – if a real effect does occur, the chance of your actually detecting it may be very low – often less than 20%, rather than the commonly used 80%. That means that there is a high probability that you'll get a non-significant result that is really a non-result – a result in which you have little confidence, and your resources will have been wasted. You may be lucky, and the effect of your treatments may be much larger than the one you aimed to detect, but that result is unlikely.

How much should you gamble? Again, there's no simple answer, as we are dealing with a continuum, rather than a clear cut-off. If the power is 75%, you wouldn't be too worried about proceeding, but what of 70%? 50%? The decision will most often be the result of a suite of considerations. How exciting would a significant result be? How important is it that we get some information, even if it's not conclusive? Will some other people add to my data, so eventually we'll be able to get a clear answer to the hypothesis? Would an unpublishable non-significant result be a career impediment? The answer to the last question depends on who you are, what stage of your career you are at, how strong your scientific record is, and so on.

If you aren't willing to gamble, you have only a couple of options. The first is to look hard at the experimental design. Are there ways to make the experiment more efficient, so I need less time or money to deal with each replicate? Decreasing the resources needed for each experimental unit may allow you to increase the sample size. Alternatively, are there other variables that could be incorporated into the design that might reduce the background noise?

The second option, which is intermediate between a calculated gamble and rethinking the analysis, is the approach described in Chapter 3, in which we don't regard the rates of Type I and Type II errors as fixed. One conventional approach would be to use a less stringent criterion for statistical significance, i.e., increase α, producing an increase in power. This solution isn't satisfactory,

as we would still be allowing the Type II error rate to fluctuate according to logistic constraints, and just fixing the Type I error rate at a new value. The solution proposed by Mapstone (1995) is that, when we must compromise an experimental design, we do so by preserving the relative sizes of the two errors. He suggests that, as part of the design phase, we have identified the desirable error rates, and those two rates should be chosen to reflect our perception of the importance of the two kinds of errors. He suggested that compromises should preserve those relative rates, so that if we proceed with a less than optimal experiment, we are more likely to make both kinds of decision errors. That approach has been detailed for environmental monitoring by Keough & Mapstone (1995, 1997), including a flow diagram to detail those authors' view of how a sampling program gets designed. This approach is sensible, but it is too soon to see if it will gain wide acceptance in the broader scientific literature.

Occasionally, the calculations may show that the MDES is much less than the desirable effect size, suggesting that the experimental/sampling program is more sensitive than expected. In this case, you could consider reducing the replication, with the possibility of using "spare" resources for further studies. Our experience suggests that this latter situation is uncommon.

While formal power analysis is part of the Neyman–Pearson approach (Chapter 3), and most often discussed as part of hypothesis testing, the general principles apply to other statistical tasks. When estimating the value of a particular parameter, we may wish to be certain that we produce an accurate estimate of that parameter (Section 7.1.2), and the confidence that we have in that estimate will be similar to power, depending on sampling effort, variability, etc. If our aim is to produce a confidence interval around an estimate, the procedures become even more similar – a confidence interval requires a statement about the level of confidence, e.g. 0.95, and depends also on sampling effort and variation. We must also make some decision about the distribution of our parameter, either by assigning a formal distribution (e.g. normal, Poisson), or by opting for a randomization procedure.

A priori power analysis should, we think, be a routine part of planning any experiment. Our initial power estimates may be quite crude, especially when we have a poor estimate of the variation present in our data. As we will see in later chapters, too, for complex designs, we may be faced with a large range of power curves, corresponding to different patterns among our treatments, and we will not be sure what pattern to expect. However, we will at least know whether "important" effects are likely to be detected, given our available resources. Having that knowledge makes us decide whether to reallocate our resources to maximize the power for our key hypotheses.

Perhaps the most valuable part of *a priori* power analysis is that, to do the calculations, we must specify the alternative hypothesis, and, most importantly, the statistical model that we will apply to the data. Specifying the model makes us think about the analysis *before* the data have been collected, a habit that we recommend strongly.

The final, important point is that power calculations, especially at the planning stage, are approximate. We usually use pilot estimates of variation that, if we do the appropriate calculations, tend to have alarmingly large confidence intervals, so our power estimates will also have considerable imprecision. If our target power value is 0.80, we should be looking for calculations that give power values in this region. Often, our sample sizes in biological work are quite small, and power values move in substantial increments, because the sample size, n, is an integer. In planning, we should not focus on whether power is 0.75, 0.80, etc., but on making sure we have enough samples to approach the desirable value, rather than giving values of 0.30 or 0.40.

7.4 General issues and hints for analysis

7.4.1 General issues

- When thinking about experimental design, the need for appropriate controls is familiar to most researchers, but less attention is often paid to appropriate units of replication. It is crucial to identify, for a particular hypothesis,

and set of experimental treatments, the experimental units to which these treatments are applied. These experimental units are the replicates for testing that hypothesis.
- In more complex designs, testing several hypotheses, the experimental units may occur at several temporal and spatial scales. Attention must be paid to identifying the appropriate amount of replication for each of these hypotheses.
- Power analysis, used when planning a sampling or experimental program, provides a means of determining whether our plan is feasible, or of deciding the resources that are necessary for a particular experiment.
- A power analysis can only be done when we have an estimate of the variation in the system under study. If the power analysis is done before sampling, we must obtain an estimate of variation on the same spatial and temporal scale as our planned experimental units.
- Power analysis also requires us to specify the statistical model that will be applied to the data – without this step, no calculations can be made. While we may be forced to make changes when the real data arrive, this step is useful in formalizing our experimental design.
- Power equations can be used to determine the number of replicates (at the planning stage), the change that could be detected (at planning or analysis stages), or the degree of confidence in the analysis (after a non-significant result).
- The most difficult task is almost always determining an important effect size, but doing so focuses our attention on what is biologically important, rather than just looking for statistical significance.

7.4.2 Hints for analysis

- At the planning stage, write out an analysis table and its associated statistical model, to be sure that you understand the design clearly. Identify the key hypothesis tests.
- Determine the effect size by thinking about what would be important biologically.
- Focus on using power analysis to determine appropriate sample sizes in the design stage. *Post hoc* power calculations can be useful for pre-specified effect sizes. Calculating observed power, the power to detect the observed effect, is pointless.
- The formal analysis of power for simple designs can now be done using a wide range of software packages.
- More complex analyses require an understanding of the calculation of non-centrality parameters. After making that calculation, non-central distribution functions are freely available for most common statistical distributions.

Chapter 8

Comparing groups or treatments – analysis of variance

The analysis of variance (ANOVA) is a general statistical technique for partitioning and analyzing the variation in a continuous response variable. We used ANOVA in Chapters 5 and 6 to partition the variation in a response variable into that explained by the linear regression with one or more continuous predictor variables and that unexplained by the regression model. In applied statistics, the term "analysis of variance" (ANOVA) is commonly used for the particular case of partitioning the variation in a response variable into that explained and that unexplained by one or more categorical predictors, called factors, usually in the context of designed experiments (Sokal & Rohlf 1995, Underwood 1997). The categories of each factor are the groups or experimental treatments and the focus is often comparing response variable means between groups. We emphasized in Chapter 5 that the statistical distinction between "classical regression" and "classical ANOVA" is artificial. Both involve the general technique of partitioning variation in a response variable (analysis of variance) and of fitting linear models to explain or predict values of the response variable. It turns out that ANOVA can also be used to test hypotheses about group (treatment) means.

The two main aims of classical ANOVA, therefore, are:

- to examine the relative contribution of different sources of variation (factors or combination of factors, i.e. the predictor variables) to the total amount of the variability in the response variable, and
- to test the null hypothesis (H_0) that population group or treatment means are equal.

8.1 Single factor (one way) designs

A single factor or one way design deals with only a single factor or predictor, although that factor will comprise several levels or groups. Designs that can be analyzed with single factor ANOVA models are completely randomized (CR) designs, where there is no restriction on the random allocation of experimental or sampling units to factor levels. Designs that involve restricted randomization will be described in Chapters 10 and 11. We will use two recent examples from the literature to illustrate use of this analysis.

Diatom communities and heavy metals in rivers
Medley & Clements (1998) studied the response of diatom communities to heavy metals, especially zinc, in streams in the Rocky Mountain region of Colorado, USA. As part of their study, they sampled a number of stations (between four and seven) on six streams known to be polluted by heavy metals. At each station, they recorded a range of physicochemical variables (pH, dissolved oxygen etc.), zinc concentration, and variables describing the diatom community (species richness, species diversity H′ and proportion of diatom cells that were the early-successional species, *Achanthes minutissima*). One of their analyses was to ignore streams and partition the 34 stations into four zinc-level categories: background (<20 $\mu g\ l^{-1}$, 8 stations), low (21–50 $\mu g\ l^{-1}$, 8 stations), medium

51–200 μg l^{-1}, 9 stations), and high (>200 μg l^{-1}, 9 stations) and test the null hypothesis that there were no differences in diatom species diversity between zinc-level groups, using stations as replicates. We will also use these data to test the null hypothesis that there are no differences in diatom species diversity between streams, again using stations as replicates. The full analyses of these data are in Box 8.1.

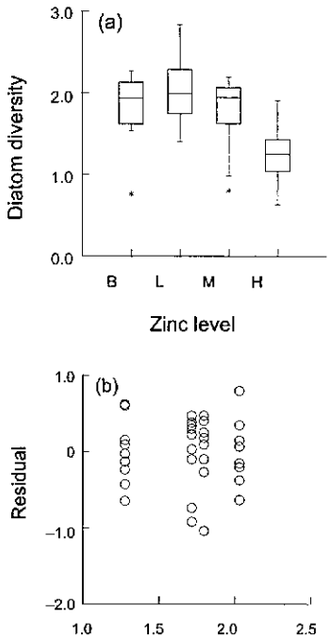

Figure 8.1 (a) Boxplots of diatom diversity against zinc-level group from Medley & Clements (1998). B is background, L is low, M is medium and H is high zinc level. (b) Residual plot from fit of single factor ANOVA model relating diatom diversity to zinc-level group from Medley & Clements (1998).

Box 8.1 Worked example: diatom communities in metal-affected streams

Medley & Clements (1998) sampled a number of stations (between four and seven) on six streams known to be polluted by heavy metals in the Rocky Mountain region of Colorado, USA. They recorded zinc concentration, and species richness and species diversity of the diatom community and proportion of diatom cells that were the early-successional species, *Achanthes minutissima*.

Species diversity versus zinc-level group

The first analysis compares mean diatom species diversity (response variable) across the four zinc-level groups (categorical predictor variable), zinc level treated as a fixed factor. The H_0 was no difference in mean diatom species diversity between zinc-level groups. Boxplots of species diversity against group (Figure 8.1(a)) showed no obvious skewness; two sites with low species diversity were highlighted in the background and medium zinc groups as possible outliers. The results from an analysis of variance from fitting a linear model with zinc level as the predictor variable were as follows.

Source	SS	df	MS	*F*	*P*
Zinc level	2.567	3	0.856	3.939	0.018
Residual	6.516	30	0.217		
Total	9.083	33			

The residual plot from this model (Figure 8.1(b)) did not reveal any outliers or any unequal spread of the residuals, suggesting the assumptions of the ANOVA were appropriate. Additionally, Levene's test produced no evidence that the H_0 of no differences in variances of species diversity between the zinc-level groups should be rejected (Levene-mean: $F_{3,30} = 0.087$, $P = 0.967$; Levene-median: $F_{3,30} = 0.020$, $P = 0.996$).

Tukey's pairwise comparison of group means: mean differences with Tukey adjusted *P* values for each pairwise comparison in brackets.

	Background	Low	Medium	High
Background	0.000 (1.000)			
Low	0.235 (0.746)	0.000 (1.000)		
Medium	0.080 (0.985)	0.315 (0.515)	0.000 (1.000)	
High	0.520 (0.122)	0.755 (0.012)	0.440 (0.209)	0.000 (1.000)

The only H_0 to be rejected is that of no difference in diatom diversity between sites with low zinc and sites with high zinc.

We could also analyze these data with more robust methods, especially if we were concerned about underlying non-normality or outliers. To test the H_0 that there is no difference in the location of the distributions of diatom diversity between zinc levels, irrespective of the shape of these distributions, we would use the Kruskal–Wallis non-parametric test based on ranks sums.

Zinc level	Rank sum
Background	160.0
Low	183.0
Medium	166.5
High	85.5

The Kruskal–Wallis *H*-statistic equals 8.737. The probability of getting this value of one more extreme when the H_0 is true (testing with a chi-square distribution with 3 df) is 0.033, so we would reject the H_0.

We might also consider a randomization test, where we reallocate observations to the four groups at random many times to generate a distribution of a suitable test statistic. We used Manly's (1997) program RT, the percentage of total SS attributable to zinc levels (groups) as the statistic and used 1000 randomizations. The percentage of SS_{Total} accounted for by SS_{Groups} was 28.3% and the probability of getting this value or one more extreme if the H_0 of no effects of zinc level on diatom diversity was true was 0.023. Again, we would reject the H_0 at the 0.05 level.

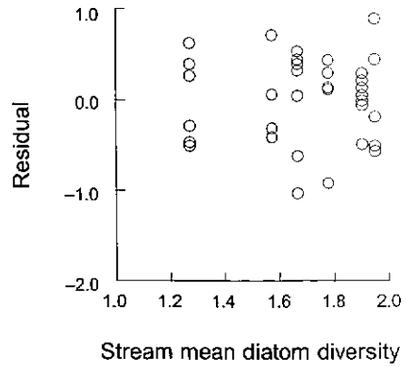

Figure 8.2 Residual plot from fit of single factor random effects ANOVA model relating diatom diversity to stream group from Medley & Clements (1998).

Species diversity versus stream

The second analysis compared diatom species diversity across the streams. Streams are treated as a random factor, assuming these streams represent a random sample of all possible streams in this part of the Rocky Mountains. The H_0 then is that there is no added variance (above the variation between stations) due to differences in diatom species diversity between streams in this part of the Rocky Mountains.

Source	SS	df	MS	*F*	*P*
Stream	1.828	5	0.366	1.411	0.251
Residual	7.255	28	0.259		
Total	9.083	33			

The residual plot (Figure 8.2) indicates no variance heterogeneity, although the sample sizes within each stream are too small for useful boxplots. We used the ANOVA, ML and REML methods to estimate the two variance components (σ_ε^2 and σ_α^2). ML and REML estimates are tedious to calculate by hand so we used SPSS (Ver 9.0) to obtain these estimates. Confidence intervals (95%) are provided for σ_ε^2 only; unequal sample sizes preclude reliable confidence intervals for σ_α^2.

Method	Estimate of σ_ε^2	Estimate of σ_α^2
ANOVA	0.259 (0.159–0.452)	0.0189
ML	0.257	0.0099
REML	0.258	0.0205

Note that there is little difference in the estimates of σ_ε^2, although both ML and REML estimates will be biased. The estimates of σ_α^2 differ considerably between estimation methods, however. Based on Section 8.2.1, the REML estimate of 0.0205 is probably the most reliable. Most of the variance is due to differences between stations within streams rather than due to differences between all possible streams.

Settlement of invertebrate larvae

Keough & Raimondi (1995) were interested in the degree to which biofilms – films of diatoms, algal spores, bacteria, and other organic material – that develop on hard surfaces influence the settlement of invertebrate larvae. In an earlier paper, from southeastern Australia, Todd & Keough (1994) had manipulated these biofilms by covering experimental surfaces with fine mesh that excluded most larvae, but allowed diatoms, etc., to pass through. These nets were then removed to allow invertebrates to settle. Keough & Raimondi focused on the ability of larvae to respond to successional changes that occur in biofilms, and, because the earlier procedure was time-consuming, decided to test whether the films that developed in laboratory seawater systems had similar effects to those developing in the field. At the same time, they tested whether covering a surface with netting altered the biofilm (or at least its attractiveness to larvae). They used four experimental treatments: substrata that had been conditioned in sterile seawater, surfaces immersed in laboratory aquaria, surfaces in laboratory aquaria, but with fine mesh netting over the surface, and surfaces immersed in the field, and covered with identical netting. After one week for biofilms to develop, the experimental surfaces (11 cm × 11 cm pieces of Perspex (Plexiglas)) were placed in the field in a completely randomized array. They were left for one week, and then the newly settled invertebrates identified and counted. To control for small numbers of larvae passing through the netting during the conditioning period, they used an additional treatment, which was netted, and returned to the laboratory after one week and censused. The values of this treatment were used to adjust the numbers in the treatment that started in the field. The data for analysis then consisted of four treatments: sterile, lab films with net, lab films without net, and field films with net. We will use their data to test the null hypothesis that there are no differences in recruitment of one family of polychaete worms, the serpulids, and to specifically compare some combinations of treatments. The analyses of these data are in Box 8.2 and Box 8.4.

8.1.1 Types of predictor variables (factors)

There are two types of categorical predictor variables in linear models. The most common type is a fixed factor, where all the levels of the factor (i.e. all the groups or treatments) that are of interest are included in the analysis. We cannot extrapolate our statistical conclusions beyond these specific levels to other groups or treatments not in the study. If we repeated the study, we would usually use the same levels of the fixed factor again. Linear models based on fixed categorical predictor variables (fixed factors) are termed fixed effects models (or Model 1 ANOVAs). Fixed effect models are analogous to linear regression models where X is assumed to be fixed. The other type of factor is a random factor, where we are only using a random selection of all the possible levels (or groups) of the factor and we usually wish to make inferences about all the possible groups from our sample of groups. If we repeated the study, we would usually take another sample of groups from the population of possible groups. Linear models based on random categorical predictor variables (random factors) are termed random effects models (or Model 2 ANOVAs). Random effects models are analogous to linear regression models where X is random (Model II regression; see Chapter 5).

To illustrate the difference between these types of factors, the zinc-level groups created by Medley & Clements (1998) clearly represent a fixed factor. These groups were specifically chosen to match the USA EPA chronic criteria values for zinc and any further study would definitely use the same groupings. Any conclusions about differences in diatom communities between zinc levels are restricted to these specific groups. In contrast, we might consider the six streams used by Medley & Clements (1998) as a possible random sample from all metal-polluted streams in the southern Rocky Mountain ecoregion of Colorado and hence treat streams as a random factor. A new study might choose a different sample of streams from this region. Conclusions from our analysis could be extrapolated to all metal-polluted streams in this region.

We argue that the random (or at least haphazard) nature of the selection of groups for a random factor is important for valid interpretation of the subsequent analysis. Selecting specific levels of a factor and then calling the factor random simply to allow extrapolation to some population of levels is inappropriate, just as would be selecting a specific set of observations from a population and calling that set a random sample.

Our conclusions for a fixed factor are restricted to those specific groups we used in the experiment or sampling program. For a random factor, we wish to draw conclusions about the population of groups from which we have randomly chosen a subset. Random factors in biology are often randomly chosen spatial units like sites or blocks. Time (e.g. months or years) is also sometimes considered a random factor but it is much more difficult to envisage a sequence of months (or years) being a random sample from a population of times to which we would wish to extrapolate.

Although the distinction between fixed and random factors does not affect the model fitting or calculations for subsequent hypothesis tests in a single factor model, the hypotheses being tested are fundamentally different for fixed and random factors. When we consider more complex experimental designs in later chapters, it will be clear that the distinction between fixed and random factors can also affect the calculation of the hypothesis tests.

Box 8.2 Worked example: serpulid recruitment onto surfaces with different biofilms

Keough & Raimondi (1995) set up an experiment to examine the response of serpulid (polychaete worms) larvae to four types of biofilms on hard substrata in shallow marine waters. The four treatments were: sterile substrata, biofilms developed in the lab with a covering net, lab biofilms without a net, and biofilms developed in the field with a net. The substrata were left for one week, and then the newly settled worms identified and counted. To control for small numbers of larvae passing through the netting during the conditioning period, they used an additional treatment, which was netted, and returned to the laboratory after one week and censused. The values of this treatment were used to adjust the numbers in the treatment that started in the field.

We have not shown the initial data screening stages, but the response variable was log-transformed to improve skewed distributions. The H_0 was that there was no difference between treatments in the mean log-transformed number of serpulid recruits per substratum. The residual plot from the single factor model 8.3 with log-transformed numbers of serpulid recruits revealed a single outlier, but very similar spread of data between groups, suggesting that the assumptions were met. The similarity of data ranges is probably a more reliable guide to the reliability of the ANOVA than the formal identification of outliers from boxplots, when there are only seven observations per group.

The results from the analysis of variance were as follows.

Source	SS	df	MS	F	P
Biofilms	0.241	3	0.080	6.006	0.003
Residual	0.321	24	0.013		
Total	0.562	27			

We would reject the H_0 of no difference between treatments in the log numbers of serpulid recruits. In this particular example, however, we are more interested in the planned contrasts between specific treatments (Box 8.4).

8.1.2 Linear model for single factor analyses

Linear effects model

We introduced linear models in Chapters 5 and 6 for regression analysis. The structure of the linear model when the predictor variable is categorical is similar to those models, although there are two types of models we can fit (Box 8.3). Consider a data set consisting of p groups or treatments ($i = 1$ to p) and n replicates ($j = 1$ to n) within each group (Figure 8.1). From Medley & Clements (1998), p equals four zinc levels and n equals eight or nine stations. From Keough & Raimondi (1995), p equals four biofilm treatments and n equals seven substrata.

The linear effects model is:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad (8.1)$$

The details of the linear single factor ANOVA model, including estimation of its parameters and means, are provided in Box 8.3 and Table 8.1. OLS means and their standard errors are standard output from linear models routines in statistical software.

From Medley & Clements (1998):

$$(\text{diatom species diversity})_{ij} = \mu + (\text{effect of zinc level})_i + _{ij} \quad (8.2)$$

From Keough & Raimondi (1995):

$$(\text{no. of serpulids})_{ij} = \mu + (\text{effect of biofilm type})_i + \varepsilon_{ij} \quad (8.3)$$

Box 8.3 Single factor ANOVA models, overparameterization and estimable functions

Consider a data set consisting of p groups or treatments ($i = 1$ to p) and n replicates ($j = 1$ to n) within each group (Figure 8.4).

The linear effects model is:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

In this model:

- y_{ij} is the jth replicate observation of the response variable from the ith group of factor A;
- μ is the overall population mean of the response variable (also termed the constant because it is constant for all observations);
- if the factor is fixed, α_i is the effect of ith group (the difference between each group mean and the overall mean $\mu_i - \mu$);
- if the factor is random, α_i represents a random variable with a mean of zero and a variance of σ_α^2, measuring the variance in mean values of the response variable across all the possible levels of the factor that could have been used;
- ε_{ij} is random or unexplained error associated with the jth replicate observation from the ith group. These error terms are assumed to be normally distributed at each factor level, with a mean of zero ($E(\varepsilon_{ij})$ equals zero) and a variance of σ_ε^2.

This model is structurally similar to the simple linear regression model described in Chapter 5. The overall mean replaces the intercept as the constant and the treatment or group effect replaces the slope as a measure of the effect of the predictor variable on the response variable. Like the regression model, model 8.1 has two components: the model ($\mu + \alpha_i$) and the error (ε_{ij}).

We can fit a linear model to data where the predictor variable is categorical in a form that is basically a multiple linear regression model with an intercept. The factor levels (groups) are converted to dummy variables (Chapter 6) and a multiple regression model is fitted of the form:

$$y_{ij} = \mu + \beta_1(\text{dummy}_1)_{ij} + \beta_2(\text{dummy}_2)_{ij} + \beta_3(\text{dummy}_3)_{ij} + \ldots + \beta_{p-1}(\text{dummy}_{p-1})_{ij} + \varepsilon_{ij}$$

Fitting this type of model is sometimes called dummy coding in statistical software. The basic results from estimation and hypothesis testing will be the same as when fitting the usual ANOVA models (effects or means models) except that estimates of group effects will often be coded to compare with a reference category so only $p - 1$ effects will be presented in output from statistical software. You should always check which category your preferred software uses as its reference group when fitting a model of this type.

The linear effects model is what statisticians call "overparameterized" (Searle 1993) because the number of group means (p) is less than the number of parameters to be estimated ($\mu, \alpha_1 \ldots \alpha_p$). Not all parameters in the effects model can be estimated by OLS unless we impose some constraints because there is no unique solution to the set of normal equations (Searle 1993). The usual constraint, sometimes called a sum-to-zero constraint (Yandell 1997), a Σ-restriction (Searle 1993), or a side condition (Maxwell & Delaney 1990), is that the sum of the group effects equals zero, i.e. $\sum_{i=1}^{p} \alpha_i = 0$. This constraint is not particularly problematical for single factor designs, although similar constraints for some multifactor designs are controversial (Chapter 9). The sum-to-zero constraint is not the only way of allowing estimation of the overall mean and each of the α_i. We can also set one of the parameters, either μ or one of the α_i, to zero (set-to-zero constraint; Yandell 1997), although this approach is only really useful when one group is clearly a control or reference group (see also effects coding for linear models in Chapter 5).

An alternative single factor ANOVA model is the cell means model. It simply replaces $\mu + \alpha_i$ with μ_i and therefore uses group means instead of group effects (differences between group means and overall mean) for the model component:

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

The cell means model is no longer overparameterized because the number of parameters in the model component is obviously the same as the number of group means. While fitting such a model makes little difference in the single factor case, and the basic ANOVA table and hypothesis tests will not change, the cell means model has some advantages in more complex designs with unequal sample sizes or completely missing cells (Milliken & Johnson 1984, Searle 1993; Chapter 9).

Some linear models statisticians (Hocking 1996, Searle 1993) regard the sum-to-zero constraint as an unnecessary complication that limits the practical and pedagogical use of the effects model and can cause much confusion in multifactor designs (Nelder & Lane 1995). The alternative approach is to focus on parameters or functions of parameters that are estimable. Estimable functions are "those functions of parameters which do not depend on the particular solution of the normal equations" (Yandell 1997, p. 111). Although all of the α_i are not estimable (at least, not without constraints), $(\mu + \alpha_i)$ is estimable for each group. If we equate the effects model with the cell means model:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} = \mu_i + \varepsilon_{ij}$$

we can see that each estimable function $(\mu+\alpha_i)$ is equivalent to the appropriate cell mean (μ_i), hence the emphasis that many statisticians place on the cell means model. In practice, it makes no difference for hypothesis testing whether we fit the cell means or effects model. The *F*-ratio statistic for testing the H_0 that $\mu_1=\mu_2=\ldots=\mu_i=\ldots=\mu$ is identical to that for testing the H_0 that all α_i equal zero.

We prefer the effects model for most analyses of experimental designs because, given the sum-to-zero constraints, it allows estimation of the effects of factors and their interactions (Chapter 9), allows combinations of continuous and categorical variables (e.g. analyses of covariance, Chapter 12) and is similar in structure to the multiple linear regression model. The basic features of the effects model for a single factor ANOVA are similar to those described for the linear regression model in Chapter 5. In particular, we must make certain assumptions about the error terms (ε_{ij}) from the model and these assumptions equally apply to the response variable.

1. For each group (factor level, *i*) used in the design, there is a population of Y-values (y_{ij}) and error terms (ε_{ij}) with a probability distribution. For interval estimation and hypothesis testing, we assume that the population of y_{ij} and therefore ε_{ij} at each factor level (i) has a normal distribution.

2. These populations of y_{ij} and therefore ε_{ij} at each factor level are assumed to have the same variance $(\sigma_\varepsilon^2$, sometimes simplified to σ^2 when there is no ambiguity). This is termed the homogeneity of variance assumption and can be formally expressed as $\sigma_1^2=\sigma_2^2=\ldots=\sigma_i^2=\ldots=\sigma_\varepsilon^2$.

3. The y_{ij} and the ε_{ij} are independent of, and therefore uncorrelated with, each other within each factor level and across factor levels if the factor is fixed or, if the factor is random, once the factor levels have been chosen (Neter *et al.* 1996).

These assumptions and their implications are examined in more detail in Section 8.3.

There are three parameters to be estimated when fitting model 8.1: μ, α_i and σ_ε^2, the latter being the variance of the error terms, assumed to be constant across factor levels. Estimation of these parameters can be based on either OLS or ML and when certain assumptions hold (see Section 8.3), the estimates for μ and α_i are the same whereas the ML estimate of σ_ε^2 is slightly biased (see also Chapter 2). We will focus on OLS estimation, although ML is important for estimation of some parameters when sample sizes differ between groups (Section 8.2).

The OLS estimates of μ, μ_i and α_i are presented in Table 8.1. Note the estimate of α_i is simply the difference between the estimates of μ_i and μ. Therefore, the predicted or fitted values of the response variable from our model are:

$$\hat{y}_{ij}=\bar{y}+(\bar{y}_i-\bar{y})=\bar{y}_i$$

So any predicted Y-value is simply predicted by the sample mean for that factor level.

In practice, we tend not to worry too much about the estimates of μ and α_i because we usually focus on estimates of group means and of differences or contrasts between group means for fixed factors (Section 8.6) and components of variance for random factors (Section 8.2). Standard errors for these group means are:

$$s_{\bar{y}_i}=\sqrt{\frac{MS_{Residual}}{n_i}}$$

and confidence intervals for μ_i can be constructed in the usual manner based on the *t* distribution.

The error terms (ε_{ij}) from the linear model can be estimated by the residuals, where a residual (e_{ij}) is simply the difference between each observed and predicted Y-value $(y_{ij}-\bar{y}_i)$. Note that the sum of the residuals within each factor level equals zero $(\Sigma_{j=1}^{n} e_{ij}=0)$. The OLS estimate of σ_ε^2 is the sample variance of these residuals and is termed the Residual (or Error) Mean Square (Table 8.1); remember from Chapter 2 that a mean square is just a variance.

Box 8.4 Worked example: planned comparisons of serpulid recruitment onto surfaces with different biofilms

The mean log number of serpulid recruits for each of the four biofilm treatments from Keough & Raimondi (1995) were (see also Figure 8.3) as follows.

Treatment	Field (F)	Netted lab (NL)	Sterile lab (SL)	Un-netted lab (UL)
Log mean number of serpulid recruits	2.117	2.185	1.939	2.136

A series of planned comparisons were done, each testing a hypothesis about the nature of the biofilms. The contrasts were done in sequence, with each comparison depending on the result of previous ones.

First, Keough & Raimondi (1995) tested whether the presence of a net over a surface affected recruitment, by comparing the netted and un-netted laboratory treatments. The H_0 is:

$$\mu_{NL}=\mu_{UL} \text{ or } \mu_{NL}-\mu_{UL}=0$$

We use the latter expression to define the linear contrast equation:

$$(0)\bar{y}_F+(+1)\bar{y}_{NL}+(0)\bar{y}_{SL}+(-1)\bar{y}_{UL}$$

Note that this contrast specifically represents the H_0 and we use coefficients of zero to omit groups that are not part of the H_0. This linear contrast can be used to calculate the SS due to this comparison. The complete ANOVA table below indicates that we would not reject this H_0.

Second, the laboratory and field films were compared. Because the two kinds of laboratory-developed films did not differ, we can pool them, so the H_0 is:

$$(\mu_{NL}+\mu_{UL})/2=\mu_F \text{ or } (\mu_{NL}+\mu_{UL})/2-\mu_F=0$$

The linear contrast equation is:

$$(+1)\bar{y}_F+(-0.5)\bar{y}_{NL}+(0)\bar{y}_{SL}+(-0.5)\bar{y}_{UL} \text{ or}$$

$$(+2)\bar{y}_F+(-1)\bar{y}_{NL}+(0)\bar{y}_{SL}+(-1)\bar{y}_{UL}$$

Note that the coefficients for the two lab treatments produce the average of those two groups, which is contrasted to the field treatment. The ANOVA table below indicates that we would not reject this H_0.

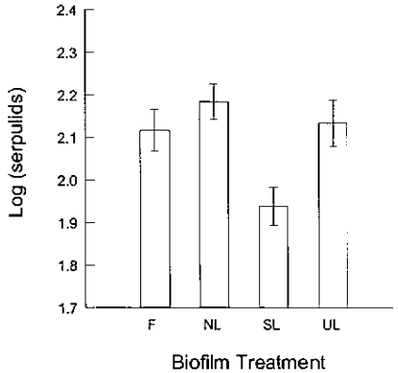

Figure 8.3 Plot of means and standard errors of log number of serpulid recruits for the four biofilm treatments used by Keough & Raimondi (1995). F denotes the treatment with biofilms developing in the field; NL, UL and SL indicates biofilms developed under laboratory conditions, with NL and UL being substrata in laboratory aquaria with (N) and without (U) nets, and SL substrata being immersed in sterile seawater.

Finally, we compare the whole set of substrata with biofilms present, to the single, unfilmed treatment. The H_0 is:

$(\mu_F + \mu_{NL} + \mu_{UL})/3 = \mu_{SL}$ or $(\mu_F + \mu_{NL} + \mu_{UL})/3 - \mu_{SL} = 0$

The linear contrast equation is:

$(+1)\bar{y}_F + (-0.33)\bar{y}_{NL} + (-0.33)\bar{y}_{SL} + (-0.33)\bar{y}_{UL}$ or

$(+3)\bar{y}_F + (-1)\bar{y}_{NL} + (-1)\bar{y}_{SL} + (-1)\bar{y}_{UL}$

Now the coefficients for the three lab treatments represent the average of those three groups and is contrasted to the field treatment. We would reject this H_0.

Source	SS	df	MS	F	P
Biofilms	0.241	3	0.080	6.006	0.003
NL vs UL	0.008	1	0.008	0.635	0.433
F vs average (NL & UL)	0.008	1	0.008	0.644	0.423
SL vs average (F & NL & UL)	0.217	1	0.217	16.719	<0.001
Linear trend	0.079	1	0.079	6.096	0.021
Residual	0.321	24	0.013		
Total	0.562	27			

Note that as long as the coefficients sum to zero (i.e. $\Sigma_{i=1}^{p} n_i c_i = 0$) and represent the contrast of interest, the size of the coefficients is irrelevant, e.g. in the first example above, we could have used 1, −1,0,0 or 0.5, −0.5,0,0 or 100, −100,0,0, the results would be identical. Note also that these comparisons are orthogonal. For example, for the first two comparisons, we can use the formal test of orthogonality $\Sigma_{i=1}^{p} c_{i1} c_{i2} = (0)(1) + (1)(-0.5) + (0)(0) + (-1)(-0.5) = 0 - 0.5 + 0 + 0.5 = 0$.

Although Keough & Raimondi did not ask this question, it could have been that the sterile water and the three biofilm treatments became monotonically richer as a cue for settlement. If so, a test for trend would have been appropriate, with the four treatments ranked SL, NL, UL, F and considered equally spaced. Using the information in Table 8.8, the contrast equation is:

$(+3)\bar{y}_F + (-1)\bar{y}_{NL} + (-3)\bar{y}_{SL} + (+1)\bar{y}_{UL}$

The results for this contrast are in the ANOVA table above and we would reject the H_0 and conclude that there is a trend, although inspection of the means (Figure 8.3) suggests that the trend is influenced by the low settlement of worms onto the unfilmed (SL) treatment. If we had decided to test for a quadratic trend, our coefficients would be of the form 1 −1 −1 1, and, in the order in which our treatments are listed, the coefficients would be 1 −1 1 −1. Such a trend is not of much interest here.

Table 8.1 Parameters, and their OLS estimates, from a single factor linear model with example calculations illustrated for diatom species diversity in different zinc-level groups from Medley & Clements (1998)

Parameter	Estimate	Medley & Clements (1998)
μ_i	$\bar{y}_i = \frac{\sum_{j=1}^{n} y_{ij}}{n_i}$	Group mean (±SE) diversity: Background 1.797 ± 0.165 Low 2.033 ± 0.165 Medium 1.718 ± 0.155 High 1.278 ± 0.155
μ	$\bar{y} = \frac{\sum_{i=1}^{p} \bar{y}_i}{p}$	Overall mean diversity: 1.694
$\alpha_i = \mu_i - \mu$	$\bar{y}_i - \bar{y}$	Background: 1.797 − 1.694 = 0.103 Low: 2.033 − 1.694 = 0.339 Medium: 1.718 − 1.694 = 0.024 High: 1.278 − 1.694 = −0.416
ε_{ij}	$e_{ij} = y_{ij} - \bar{y}_i$	Background: Obs 1: 2.270 − 1.797 = 0.473 Obs 2: 2.200 − 1.797 = 0.403 Obs 3: 1.890 − 1.797 = 0.093 Obs 4: 1.530 − 1.797 = −0.267 etc.
σ_ε^2	$MS_{Residual}$ $\frac{\sum_{i=1}^{p}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2}{\sum_{i=1}^{n} n_i - p}$	$= [(0.473)^2 + (0.403)^2 + (0.093)^2 + \cdots]/[(8 + 8 + 9 + 9) - 4]$

In models 8.1 and 8.2 we have the following.

y_{ij} is the jth replicate observation from the ith group, e.g. the diatom species diversity in the jth station from the ith zinc-level group.

μ is the overall population mean diatom species diversity across all possible stations from the four zinc-level groups.

If the factor is fixed, α_i is the effect of ith group (the difference between each group mean and the overall mean $\mu_i - \mu$), e.g. the effect of the ith zinc-level group on diatom species diversity, measured as the difference between the mean species diversity for the ith zinc level and the overall mean species diversity. If the factor is random, α_i represents a random variable with a mean of zero and a variance of σ_α^2, e.g. the variance in the mean diatom species diversity per stream across all the possible streams in the southern Rocky Mountains that Medley & Clements (1998) could have used in their study.

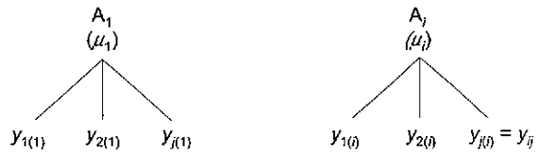

Figure 8.4 General data layout for single factor ANOVA where factor A has p ($i = 1$ to p) groups and there are n ($j = 1$ to n) replicates.

ε_{ij} is random or unexplained error associated with the jth replicate observation from the ith group. For example, this measures

Table 8.2 | ANOVA table for single factor linear model showing partitioning of variation

Source of	SS	df	MS
Between groups	$\sum_{i=1}^{p} n_i(\bar{y}_i - \bar{y})^2$	$p-1$	$\frac{\sum_{i=1}^{p} n_i(\bar{y}_i - \bar{y})^2}{p-1}$
Residual	$\sum_{i=1}^{p}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2$	$\sum_{i=1}^{p} n_i - p$	$\frac{\sum_{i=1}^{p}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2}{\sum_{i=1}^{n} n_i - p}$
Total	$\sum_{i=1}^{p}\sum_{j=1}^{n}(y_{ij} - \bar{y})^2$	$\sum_{i=1}^{p} n_i - 1$	

the error associated with each replicate observation of diatom species diversity at any possible station within any of the four zinc levels.

For interval estimation and tests of hypotheses about model parameters, we must make certain assumptions about the error terms (ε_{ij}) from model 8.1 and these assumptions also apply to the response variable. First, the population of y_{ij} and therefore ε_{ij} at each factor level (i) has a normal distribution. We assume that there is a population of stations with normally distributed diatom species diversity for each zinc level. Second, these populations of y_{ij} and therefore ε_{ij} at each factor level are assumed to have the same variance (σ_ε^2, sometimes simplified to σ^2 when there is no ambiguity). We assume the variances in diatom species diversity among stations for each zinc level are equal. Finally, the y_{ij} and the ε_{ij} are independent of, and therefore uncorrelated with, each other within each factor level and across factor levels if the factor is fixed or, if the factor is random, once the factor levels have been chosen (Neter *et al.* 1996). In the study of Medley & Clements (1998), some stations were on the same stream so what happens upstream might influence what happens downstream, an issue of concern for all stream ecologists (see Downes *et al.* 2001). We will examine these assumptions and their implications in more detail in Section 8.3.

Predicted values and residuals

The Y-values predicted from the linear model are simply the sample means for the factor level containing the observed value:

$$\hat{y}_{ij} = \bar{y}_i \qquad (8.4)$$

The error terms (ε_{ij}) from the linear model can be estimated by the residuals, where each residual (e_{ij}) is the difference between the observed and the predicted Y-value:

$$e_{ij} = y_{ij} - \bar{y}_i \qquad (8.5)$$

For example, the residuals from the model relating diatom species diversity to zinc-level group are the differences between the observed species diversity at each station and the mean species diversity for the zinc level that station came from. As in regression analysis, residuals provide the basis of the OLS estimate of σ_ε^2 and they are valuable diagnostic tools for checking assumptions and fit of our model (Section 8.4).

8.1.3 Analysis of variance

As described in Chapters 5 and 6 for regression models, ANOVA partitions the total variation in the response variable into its components or sources. This partitioning of variation is expressed in the form of an ANOVA table (Table 8.2). We first describe the variation in Y as sums of squares (SS). The SS_{Total} for Y is the sum of the squared differences between each y_{ij} and the overall mean $\bar{y}$. The degrees of freedom (df) is the

Table 8.3 | Imaginary data based on Medley & Clements (1998) showing diatom species diversity at eight stream stations in each of four zinc levels (background, low, medium, high) – see text for details. In (a), all the variation is residual and SS_{Groups} explains none of the variation in Y (no difference between group means). In (b), there is no residual variation and SS_{Groups} explains all the variation in Y

(a) Zinc level	B	L	M	H
	0.8	0.7	1.8	2.6
	0.9	1.7	2.1	0.6
	2.4	1.0	0.6	1.2
	1.4	1.4	1.1	1.3
	1.3	1.2	2.4	2.2
	1.8	2.4	1.2	0.9
	2.1	1.1	0.9	1.9
	1.0	2.2	1.6	1.0
Means	1.4625	1.4625	1.4625	1.4625
(b) Zinc level	B	L	M	H
	1.2	2.3	1.8	0.7
	1.2	2.3	1.8	0.7
	1.2	2.3	1.8	0.7
	1.2	2.3	1.8	0.7
	1.2	2.3	1.8	0.7
	1.2	2.3	1.8	0.7
	1.2	2.3	1.8	0.7
	1.2	2.3	1.8	0.7
Means	1.2	2.3	1.8	0.7

total number of observations across all groups minus one. SS_{Total} can be partitioned into two additive components.

First is the variation due to the difference between group means, calculated as the difference between each $\bar{y}_i$ and the overall mean $\bar{y}$. This is a measure of how different the group means are and how much of the total variation in Y is explained by the difference between groups, or in an experimental context, the effect of the treatments. The df associated with the variation between group means is the number of groups minus one.

Second is the variation due to the difference between the observations within each group, calculated as the difference between each y_{ij} and relevant group mean $\bar{y}_i$. This is a measure of how different the observations are within each group, summed across groups, and also how much of the total variation in Y is not explained by the difference between groups or treatments. The df associated with the $SS_{Residual}$ is the number of observations in each group minus one, summed across groups, which is equal to the sum of the sample sizes minus the number of groups.

These SS and df are additive:

$$SS_{Total} = SS_{Groups} + SS_{Residual}$$
$$df_{Total} = df_{Groups} + df_{Residual}$$

As pointed out in Chapter 5, the sum-of-squares (SS) is a measure of variation that is dependent on the number of observations that contribute to it. In contrast to the SS, the variance (mean square) is a measure of variability that does not depend on sample size because it is an average of the squared deviations (Chapter 2). We convert the SS into Mean Squares (MS) by dividing them by their df (Table 8.2).

A detailed description of the algebra behind this partitioning of the variation can be found in Underwood (1997). Note that there are re-expressions of the formulae in Table 8.2 that are much easier to use when doing the calculations by hand (Sokal & Rohlf 1995, Underwood 1981, 1997). In practice, however, statistical software will calculate the SS and MS by fitting and comparing linear models (Section 8.1.5).

The best way to appreciate the variation between groups and the residual variation is to look at two extreme imaginary data sets, based on species diversity of stream diatoms at different zinc levels (Medley & Clements 1998). The data in Table 8.3(a) show a situation in which all the variation is between observations within each group (residual) with no variation between groups (identical group means). In contrast, the data in Table 8.3(b) are where all the variation is between groups with no residual variation (all observations within each group are identical).

The mean squares from the ANOVA are sample variances and, as such, they estimate population parameters. Statisticians have determined the expected values of MS_{Groups} and $MS_{Residual}$, termed expected mean squares (EMS), i.e. the means of the probability distributions of these sample

Table 8.4 Expected mean squares for a single factor ANOVA

Source	Fixed factor (Model 1)	Random factor (Model 2)	*F*-ratio
MS_{Groups}	$\sigma_\varepsilon^2 + \sum_{i=1}^{p} n_i \frac{(\alpha_i)^2}{p-1}$	$\sigma_\varepsilon^2 + \frac{\left[\left(\sum_{i=1}^{p} n_i\right)^2 - \sum_{i=1}^{p} n_i^2\right]\sigma_\alpha^2}{\sum_{i=1}^{p} n_i (p-1)}$	$\frac{MS_{Groups}}{MS_{Residual}}$
	If equal n: $\sigma_\varepsilon^2 + n\sum_{i=1}^{p} \frac{(\alpha_i)^2}{p-1}$	If equal n: $\sigma_\varepsilon^2 + n\sigma_\alpha^2$	
$MS_{Residual}$	σ_ε^2	σ_ε^2	

variances or what population values these mean squares actually estimate (Table 8.4; see Underwood 1997 for a clear, biologically orientated, explanation).

The $MS_{Residual}$ estimates σ_ε^2, the pooled population variance of the error terms, and hence of the *Y*-values, within groups. Note that we must assume homogeneity of error variances across groups (homogeneity of variances assumption; see Sections 8.1.2 and 8.3) for this expectation to hold.

The MS_{Groups} estimates the pooled variance of the error terms across groups plus a component representing the squared effects of the chosen groups if the factor is fixed, or the variance between all possible groups if the factor is random (Table 8.4). Note that these EMS are subject to the important constraint that $\Sigma_{i=1}^{p} \alpha_i$ equals zero, i.e. the sum of the group effects equals zero. Without this constraint, we cannot get unbiased estimators of individual treatment effects (Box 8.3; Underwood 1997, Winer *et al.* 1991).

8.1.4 Null hypotheses

The null hypothesis tested in a single factor fixed effects ANOVA is usually one of no difference between group means:

H_0: $\mu_1 = \mu_2 = \ldots = \mu_i = \ldots = \mu$

We defined group effects (α_i) in Section 8.1.2 and Box 8.3 as $\mu_i - \mu$, the difference between the population mean of group i and the overall mean. This is a measure of the effect of the *i*th group, or in an experimental context, the *i*th treatment. The null hypothesis can therefore also be expressed as no effects of groups or treatments, i.e. all treatment or group effects equal zero:

H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_i = \ldots = 0$

For a random effects ANOVA, the null hypothesis is that the variance between all possible groups equals zero:

H_0: $\sigma_\alpha^2 = 0$

The EMS from our ANOVA table (Table 8.4) allow us to determine *F*-ratios for testing these null hypotheses.

If the H_0 for a fixed factor is true, all α_i equal zero (no group effects). Therefore, both MS_{Groups} and $MS_{Residual}$ estimate σ_ε^2 and their ratio should be one. The ratio of two variances (or mean squares) is called an *F*-ratio (Chapter 2). If the H_0 is false, then at least one α_i will be different from zero. Therefore, MS_{Groups} has a larger expected value than $MS_{Residual}$ and their *F*-ratio will be greater than one. A central *F* distribution is a probability distribution of the *F*-ratio when the two sample variances come from populations with the same expected values. There are different central *F* distributions depending on the df of the two sample variances (see Figure 1.2). Therefore, we can use the appropriate probability distribution of *F* (defined by numerator and denominator df) to determine whether the probability of obtaining our sample *F*-ratio or one more extreme (the usual hypothesis testing logic; see Chapter 3), is less than some specified significance level (e.g. 0.05) and therefore whether we reject H_0 or not.

If the H_0 for a random factor is true, then σ_α^2 equals zero (no added variance due to groups) and both MS_{Groups} and $MS_{Residual}$ estimate σ_ε^2 so their *F*-ratio should be one. If the H_0 is false, then σ_α^2 will be greater than zero, MS_{Groups} will have a larger expected value than $MS_{Residual}$ and their *F*-ratio will be greater than one.

These *F*-ratio tests (usually abbreviated to *F* tests) of null hypotheses for fixed and random factors are illustrated for our worked examples in Box 8.1 and Box 8.2. The construction of the tests of null hypotheses is identical for fixed and random factors in the single factor ANOVA model, but these null hypotheses have very different interpretations. The H_0 for the fixed factor refers only to the groups used in the study whereas the H_0 for the random factor refers to all the possible groups that could have been used. It should also be clear now why the assumption of equal within group variances is so important. If σ_1^2 does not equal σ_2^2, etc., then $MS_{Residual}$ does not estimate a single population variance (σ_ε^2), and we cannot construct a reliable *F*-ratio for testing the H_0 of no group effects.

8.1.5 Comparing ANOVA models

The logic of fitting ANOVA models is the same as described in Chapters 5 and 6 for linear regression models. Either OLS or ML can be used, the fit being determined by explained variance or log-likelihoods respectively. We will use OLS in this chapter.

The full effects model containing all parameters is:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad (8.1)$$

The reduced model when H_0 that all α_i equal zero is true is:

$$y_{ij} = \mu + \varepsilon_{ij} \quad (8.6)$$

Model 8.6 simply states that if there are no group effects, our best prediction for each observation is the overall mean, e.g. if there are no effects of zinc level on diatom species diversity, then our best predictor of species diversity at each station is the overall species diversity across all stations. The residual variation when the full model is fitted is the $SS_{Residual}$ from the ANOVA in Table 8.2. The residual variation when the reduced model is fitted is the SS_{Total} from the ANOVA in Table 8.2. The difference between the unexplained variation of the full model ($SS_{Residual}$) and the unexplained variation from the reduced model (SS_{Total}) is simply the SS_{Groups}. It measures how much more of the variation in *Y* is explained by the full model compared to the reduced model. It is, therefore, the relative magnitude of the MS_{Groups} that we use to evaluate the H_0 that there are no group effects. Although comparing full and reduced models is trivial for a single factor ANOVA, the model comparison approach has broad applicability for testing null hypotheses about particular parameters in more complex linear and generalized linear models.

8.1.6 Unequal sample sizes (unbalanced designs)

Unequal sample sizes within each group do not cause any computational difficulties, particularly when the ANOVA is considered as a general linear model as we have described. However, unequal sample sizes can cause other problems. First, the different group means will be estimated with different levels of precision and this can make interpretation difficult (Underwood 1997). Note that sample size is only one contributor to the precision of an estimate and some statisticians have suggested that experiments should be designed with different sample sizes depending on the inherent variability of the variable in each group or the relative importance of each group (Mead 1988). Second, the ANOVA *F* test is much less robust to violations of assumptions, particularly homogeneity of variances, if sample sizes differ (Section 8.3). The worst case is when larger variances are associated with smaller sample sizes. This is a very important reason to design experiments and sampling programs with equal sample sizes where possible. Third, estimation of group effects, particularly variance components, is much more difficult (see Section 8.2). Finally, power calculations for random effects models are difficult because when σ_ε^2 is greater than zero and sample sizes are unequal, then the *F*-ratio

$MS_{Groups}/MS_{Residual}$ does not follow an F distribution (Searle *et al.* 1992). For testing null hypotheses with fixed effects, unequal sample sizes are only really a worry if the analysis produces a result that is close to the critical level; you will not be confident enough that the P value is accurate for you to be comfortable interpreting the result of your analysis. If, however, the result is far from significant, or highly significant (e.g. $P < 0.001$), you may still be confident in your conclusions.

One solution to unequal sample sizes in single factor designs is deleting observations until all groups have the same n. We regard this practice as unnecessarily extreme; the linear models approach can deal with unequal sample sizes and biological studies often suffer from lack of power and deleting observations will exacerbate the situation, particularly if one group has a much lower sample size than others. Alternatively, we can substitute group means to replace missing observations. In such circumstances, the $df_{Residual}$ should be reduced by the number of substituted observations (Chapter 4). However, if there is no evidence that the assumption of homogeneity of variance is seriously compromised, and the difference in sample sizes is not large (which is usually the case as unequal sample sizes are often caused by one or two observations going missing), then we recommend simply fitting the linear ANOVA model. Nonetheless, we support the recommendation of Underwood (1997) that experimental and sampling programs in biology with unequal sample sizes should be avoided, at least by design.

8.2 Factor effects

In linear regression, we could measure the "effect" of the predictor variable on the response variable in a number of ways, including the standardized regression slope or the proportion of variation in Y explained by the linear regression with X (e.g. r^2). These are measures of effect size (the "effects" of X on Y), although linear regression models are often fitted to non-experimental data so we are not implying any cause–effect relationship. In designs where the predictor variable is categorical, measuring effect size is of much more interest. One measure of group or treatment effects is the variance associated with the groups over and above the residual variance. The proportion of total variance in the population(s) explained by the groups then can be expressed as (Smith 1982):

$$\eta^2 = \frac{\sigma_Y^2 - \sigma_\varepsilon^2}{\sigma_Y^2} = \frac{\sigma_\alpha^2}{\sigma_\varepsilon^2 + \sigma_\alpha^2} \quad (8.7)$$

where σ_ε^2 is the residual variance, σ_α^2 is the variance explained by the groups and σ_Y^2 is the total variance in the response variable. Our aim, then, is to estimate the parameter η^2. Petraitis (1998) termed indices like η^2 PEV (proportion of explained variance) measures. One measure of η^2 is r^2, defined here, as for linear regression models (Chapters 5 and 6), as the proportion of the total SS explained by the predictor variable (groups), i.e. SS_{Groups} / SS_{Total}. Unfortunately, r^2 is dependent on the sample size in each group and also tends to overestimate the true proportion of the total variance explained by group effects (Maxwell & Delaney 1990). We need to consider other PEV measures, noting that their calculation and interpretation depend on whether we are talking about fixed or random factors.

8.2.1 Random effects: variance components

Let's first look at random effects models because they are straightforward, at least when sample sizes are equal. In the random effects model, there are two components of variance (termed "variance components") of interest (Table 8.5). The true variance between replicate observations within each group, averaged across groups, is σ_ε^2 and is estimated by $MS_{Residual}$. The true variance between the means of all the possible groups we could have used in our study is σ_α^2 and is termed the added variance component due to groups. We can estimate this added variance explained by groups by equating the observed and expected values of the mean squares (Brown & Mosteller 1991, Searle *et al.* 1992; see Table 8.4 and Table 8.5). This method of estimating variance components is termed the ANOVA or expected mean square (EMS) method (also method of moments). There are no distributional assumptions underlying these point estimates unless confidence intervals are developed or null hypotheses tested.

Confidence intervals can be calculated for these variance components (Table 8.5; Brown &

Table 8.5 ANOVA estimates of variance components and confidence intervals for a single factor random effects model. For 95% confidence intervals, we use critical values of the χ^2 and F distributions at 0.975 for upper confidence intervals and 0.025 for lower confidence intervals (covers range of 0.95)

Variance component	ANOVA estimate	Confidence interval
σ_α^2	Unequal n: $\frac{MS_{Groups} - MS_{Residual}}{\left(\sum n_i - \sum n_i^2 / \sum n_i\right)/(p-1)}$ Equal n: $\frac{MS_{Groups} - MS_{Residual}}{n}$	Approximate for equal n only: $\pm\frac{SS_{Groups}(1 - F_{p-1,p(n-1)}/F)}{n\chi^2_{p-1}}$ where: F is F-ratio from ANOVA $F_{p-1,p(n-1)}$ is value from F distribution with $p-1$ and $p(n-1)$ df χ^2 is value from χ^2 distribution with $p-1$ df
σ_ε^2	$MS_{Residual}$	$\pm\frac{SS_{Residual}}{\chi^2}$ where: χ^2 is value from χ^2 distribution with $\sum_{i=1}^{p} n_i - p$ df
$\rho_I = \frac{\sigma_\alpha^2}{\sigma_\varepsilon^2 + \sigma_\alpha^2}$	Equal n: $\frac{MS_{Groups} - MS_{Residual}}{MS_{Groups} + (n-1)MS_{Residual}}$	Equal n: $\pm\frac{F/F_{p-1,p(n-1)} - 1}{n + F/F_{p-1,p(n-1)} - 1}$ where: F and $F_{p-1,p(n-1)}$ as defined above

Mosteller 1991, Burdick & Graybill 1992, Searle *et al.* 1992). The confidence intervals are based on the χ^2 distribution, or equivalently the F distribution ($\chi^2_{\alpha;df_1} = df_1 F_{\alpha;df_1,\infty}$), because variances are distributed as a chi-square. For 95% confidence intervals, we use critical values of the χ^2 or F distribution at 0.975 for upper confidence intervals and 0.0125 for lower confidence intervals (covers range of 0.95). These confidence intervals are interpreted as a 95% probability under repeated sampling that this interval includes the true population variance explained by the groups. Note that the confidence interval for σ_α^2 is only an approximation (Searle *et al.* 1992), although exact confidence intervals can be determined for various ratios of σ_α^2 to σ_ε^2 and $\sigma_\alpha^2 + \sigma_\varepsilon^2$. With unbalanced data, a confidence interval for σ_α^2 based on the ANOVA method is not possible, although approximations are again available but tedious to calculate (Burdick & Graybill 1992, Searle *et al.* 1992).

Note that sometimes, MS_{Groups} will be less than $MS_{Residual}$ (and the F-ratio will be less than one), resulting in a negative estimate for σ_α^2. This is a problem, because variances obviously cannot be negative, by definition. The usual recommendation is to convert a negative variance component estimate to zero (Brown & Mosteller 1991). Hocking (1993) and Searle *et al.* (1992) argued that negative variance components suggest an inappropriate model has been applied or there may be serious outliers in the data and therefore that negative variance components might be a useful diagnostic tool.

An alternative approach is to use a method of variance component estimation that specifically excludes negative estimates (Searle *et al.* 1992). This

is particularly important in multifactor unbalanced designs (Chapter 9; see also Underwood 1997). These alternatives include the following.

- Maximum likelihood estimation (MLE) that involves deriving ML equations and their iterative solutions, although the estimators are biased (remember from Chapter 2 that ML estimators for variances are biased).
- Restricted maximum likelihood estimation (REML) that is a modification of MLE that excludes μ (the only fixed parameter in the random effects model) from the likelihood function, partly to correct the bias in ML estimates.
- Minimum norm quadratic unbiased estimation (MINQUE), a method that requires solving linear equations and *a priori* "guesses" of the components to be used in the estimation procedure.

REML produces the same estimates as the ANOVA method for σ_α^2 in balanced designs whereas the ML estimate will be slightly biased. Both ML and REML also preclude negative variance estimates for σ_α^2. However, in contrast to the ANOVA method, likelihood methods must assume normality for point estimates and all methods assume normality for interval estimates and hypothesis testing. Searle *et al.* (1992) summarized the merits of the different methods for unbalanced data in the single factor model and recommended REML for estimating the added variance components due to groups (σ_α^2) and the ANOVA method for estimating the residual variance (σ_ε^2).

Ideally, estimated variance components should be provided with confidence intervals. It might be tempting for biologists working with few df to talk about an "important" factor effect based on a large estimated variance component despite a non-significant F statistic and P value. However, the confidence interval associated with the variance component is likely to include zero under such circumstances. Interpretation of variance components should only follow a rejection of the H_0 of no added variance.

To calculate the proportion of total variance due to the random factor, we simply substitute our estimators into $\sigma_\alpha^2/(\sigma_\varepsilon^2 + \sigma_\alpha^2)$, which is sometimes called the intraclass correlation (ρ_I).

8.2.2 Fixed effects

Now let's look at PEV measures for fixed factors, which are more problematical. In many cases, the effects are displayed most simply using the means for each group, but we may be interested in describing or estimating the pattern of effects across all groups. For a fixed factor, the effect of any group is α_i, the difference between that group mean and the overall mean $\mu_i - \mu$ and we can calculate the variance of these group effects $[\sum_{i=1}^{p} \alpha_i^2/(p-1)]$. This measures the true variance among the fixed population group means in the specific populations from which we have sampled. Brown & Mosteller (1991) pointed out that it is somewhat arbitrary whether we use p or $p-1$ in the denominator for this variance, although since we have used the entire population of groups (a fixed factor), dividing by p (the number of groups) may actually be more appropriate. If we use $p-1$, the estimate of this variance is identical to the estimate of the added variance component for a random effect (Table 8.5), although its interpretation is different.

Petraitis (1998) has discussed the limitations of trying to calculate the proportion of total variance in the response variable that is explained by the fixed groups. One approach (see Hays 1994) is termed omega squared (ω^2) and is the variance of the fixed group means (using p in the denominator) as a proportion of this variance plus the residual variance (Table 8.6). If we base the estimate of ω^2 in Table 8.6 with $p-1$ instead of p in the denominator, we end up with the proportion of total variance due to a random factor (the intraclass correlation). So the main computational difference between the PEV based on ω^2 (fixed factor) or ρ_I (random factor) is whether p or $p-1$ is used in the denominator for the variance between groups.

Another measure of group effects for a fixed factor was provided by Cohen (1988), based on his work on power analyses. He defined effect size (f) as the difference among means measured in units of the standard deviation between replicates within groups (Table 8.6; see also Kirk 1995). The formula looks complex but is basically measuring the ratio of the standard deviation between group means and the standard deviation between replicates within each group (Cohen 1992). In this

Table 8.6 Measures of explained group (or treatment) variance in a single factor fixed effects model illustrated for diatom species diversity in different zinc-level groups from Medley & Clements (1998)

Measure	Formula	Medley & Clements (1998)
Omega squared (ω^2)	$\frac{SS_{Groups} - (p-1)MS_{Residual}}{SS_{Total} + MS_{Residual}}$	$\frac{2.567 - (4-1)0.217}{9.083 + 0.217} = 0.206$
Cohen's effect size (f)	$\sqrt{\frac{\frac{p-1}{\sum_{i=1}^{p} n_i}(MS_{Groups} - MS_{Residual})}{MS_{Residual}}}$	$\sqrt{\frac{\frac{3}{34}(0.856 - 0.217)}{0.217}} = 0.509$

context, we are measuring the effect in the observed data. Cohen's effect size is more commonly used to set effect sizes, based on the alternative hypothesis, in power calculations (Section 8.9, Box 8.5). Note that ω^2 equals $f^2/(1+f^2)$(Petraitis 1998).

Glass & Hakstian (1969), Underwood & Petraitis (1993) and Underwood (1997) have criticized measures of variance explained for fixed factors. They argued that the population "variance" of a set of fixed groups makes no sense and this measure cannot be compared to the average population variance between observations within groups, which is a true variance (see also Smith 1982). For instance, confidence intervals around estimates of explained between groups variance are silly for fixed factors because the sampling distribution would always be based on the same fixed groups. Also, these measures of proportion of variance explained by fixed groups are difficult to compare between different analyses. However, like Smith (1982), we recommend that PEV measures are useful descriptive summaries of explained variance for fixed factor ANOVA models, and recommend that using the method of equating mean squares to their expected values provides the simplest measure that is computationally equivalent to the variance component for a random effects model. We will discuss the issue of measuring explained variance for fixed and random factors in the context of multifactor ANOVA (Chapter 9).

It is important to realize that the interpretation of a fixed treatment variance and an added variance component for a random factor is very different (Underwood 1997). The former is an estimate of the variance between these particular group means in the specific population(s) being sampled. It is not an estimate of a variance of a larger population of groups. In contrast, the variance component for a random factor estimates the variance between the means of all the possible groups that could have been used in the analysis; this variance is due to the random sample of groups chosen and represents a real variance.

8.3 Assumptions

The assumptions for interval estimation and hypothesis testing based on the single factor ANOVA model actually concern the residual or error terms (ε_{ij}) but can be equivalently expressed in terms of the response variable Y. These assumptions are similar to those for linear regression models (Chapters 5 and 6). Most textbooks state that the single factor ANOVA is robust to these assumptions, i.e. the F test and interval estimates of effects are reliable even if the assumptions are not met. However, this robustness is very difficult to quantify and is also very dependent on balanced sample sizes. The F test can become very unreliable when unequal sample sizes are combined with non-normal data with heterogeneous variances. We strongly recommend that the assumptions of the ANOVA be carefully checked before proceeding with the analysis.

Box 8.5 Variation in formal implementations of power analysis

There are two possible sources for confusion when calculating power. First, effect sizes can be expressed differently. In some cases, the effect is described as the pattern of means, or, in the case of fixed effects, the α_i values. Other authors, e.g. Cohen (1988), combine the variation among means and an estimate of σ_ε^2, to produce a standardized effect.

For example, for a two sample t test, Cohen's effect size parameter is $d = (\mu_1 - \mu_2)/\sigma_\varepsilon$, and for a one factor ANOVA, his parameter f, is given by

$$f = \sqrt{\frac{\sum_{i=1}^{p} \alpha_i^2/p}{\sigma_\varepsilon^2}}$$

which can then be estimated from the α_i values specified by the alternative hypothesis, and an estimate of residual variance.

In a similar way, the non-centrality parameter is most often expressed as λ, as defined in Equation 8.9. However, Searle (1971) defines the non-centrality parameter as $\lambda/2$, and sometimes non-centrality is defined as $\sqrt{\lambda/(p-1)}$, or as $\varphi = \sqrt{(\lambda/p)}$, with p being the number of groups.

If power is to be calculated using tabulated values, we find that most authors provide power values tabulated against φ (e.g. Kirk 1995, Winer *et al.* 1991), although Cohen (1988) provides very extensive tables of power against f and n. Note that $f = \varphi/\sqrt{n}$ This reflects a difference in philosophy, with the use of φ representing a standardization using the standard error of the mean, and f a standardization using σ_ε.

These different formulations are mathematically equivalent, but it is confusing initially to encounter different definitions of ostensibly the same parameter. It is essential that you check the formulation used by a particular author or piece of software. A good check is to use a standard example from one of the major texts, and run it through the new calculations. When the same answer is obtained, begin your own calculations.

8.3.1 Normality

We assume that the error terms, and the observations, within each group come from normally distributed populations, i.e. the ε_{ij}s are normally distributed within each group. If sample sizes and variances are similar, then the ANOVA tests are very robust to this assumption. Check for outliers, skewness and bimodality. We can check the normality assumption in a number of ways because we have replicate observations within each group (Chapter 4). Boxplots of observations or residuals (Section 8.4) within groups should be symmetrical. Means and variances from a normal distribution are independent so a plot of sample means against sample variances should show no relationship. Samples from skewed distributions will show a positive relationship between means and variances. Probability plots of the residuals are also informative (Chapter 4). There are some formal tests of normality (e.g. the Wilks & Shapiro tests; goodness-of-fit tests such as Kolmogorov–Smirnov), but we find graphical methods much more informative for checking assumptions before a linear models analysis (Chapter 4).

Because lack of normality is really only a serious problem for ANOVA models when it results in variance heterogeneity, we will consider solutions in Section 8.3.2.

8.3.2 Variance homogeneity

A very important assumption is that the variances of the error terms (and of the observations in the populations being sampled) should be approximately equal in each group. This is termed the assumption of homogeneity of variances. This is a more serious assumption than that of normality; unequal variances can seriously affect the ANOVA F test (reviewed by Coombs *et al.* 1996). Wilcox *et al.* (1986) showed by simulation that with four groups and n equal to eleven, a 4:1 ratio of largest to smallest standard deviation (i.e. a 16:1 ratio of variances) resulted in a Type I error rate of 0.109 for a nominal α of 0.05. With sample sizes of six, ten, 16 and 40 and the same standard deviation ratio (largest standard deviation associated with smallest sample size), the Type I error rate could reach 0.275. The situation is similar to that described for t tests in Chapter 3, where larger variances associated with smaller sample sizes result in increased Type I error rates and larger variances associated with larger sample sizes result in reduced power (Coombs *et al.* 1996). Unequal variances, particularly when associated with unequal sample sizes, can therefore be a problem for hypothesis tests in linear ANOVA models.

There are a number of useful checks of the homogeneity of variance assumption. Boxplots of observations within each group should have similar spread. The spread of residuals (see Section 8.4) should be similar when plotted against group means. There are formal tests of homogeneity of variance that test the H_0 that population variances are the same across groups (e.g. Bartlett's, Hartley's, Cochran's, Levene's tests; see Neter *et al.* 1996, Sokal & Rohlf 1995, Underwood 1997). We will discuss these in Section 8.8 when the research hypothesis of interest concerns group variances rather than group means. However, we suggest that such tests should not be used by themselves as preliminary checks before fitting an ANOVA model for three reasons. First, some of them are very sensitive to non-normality, especially positive skewness (Conover *et al.* 1981, Rivest 1986), a common trait of continuous biological variables. Second, we really want to know if the variances are similar enough for the ANOVA F test to still be reliable. Tests for homogeneity of variance simply test whether sample groups come from populations with equal variances. If the sample size is large, these tests could reject the H_0 of equal variances when the ANOVA F test would still be reliable. Conversely, and more importantly, if sample sizes are small, they might not reject the H_0 of equal variances when the ANOVA F test would be in trouble. Finally, tests of homogeneity of variances provide little information on the underlying cause of heterogeneous variances, and diagnostic techniques (e.g. residual plots) are still required to decide what corrective action is appropriate.

There are a number of solutions to variance heterogeneity when fitting ANOVA models. If the heterogeneity is related to an underlying positively skewed distribution of the response variable, and hence the error terms from the ANOVA model, then transformations of the response variable will be particularly useful (see Chapter 4). Alternatively, fitting generalized linear models that allow different distributions for model error terms (Chapter 13) can be effective for linear models with categorical predictors. Weighted least squares, as described for linear regression models in Chapter 5, can also be used and various robust test statistics have been developed for testing hypotheses about means when variances are unequal (see Section 8.5.1).

8.3.3 Independence

The error terms and the observations should be independent, i.e. each experimental or sampling unit is independent of each other experimental unit, both within and between groups. Underwood (1997) has provided a detailed examination of this assumption in the context of ANOVA in biology. He distinguished different types of non-independence (see also Kenny & Judd 1986).

- Positive correlation between replicates within groups, which results in an underestimation of the true σ_ε^2 and increased rate of Type I errors. Such correlation can be due, for example, to experimental organisms responding positively to each other (Underwood 1997) or sequential recording of experimental units through time (Edwards 1993).

Table 8.7 | Residuals from single factor ANOVA model

Residual	$y_{ij} - \bar{y}_i$
Studentized residual	$\dfrac{y_{ij} - \bar{y}_i}{\sqrt{\dfrac{MS_{Residual}(n_i - 1)}{n_i}}}$

- Negative correlation between replicates within groups, which results in an overestimation of σ_ε^2 and increased rate of Type II errors.

Lack of independence can also occur between groups if the response of experimental units in one treatment group influences the response in other groups. For example, an experimental treatment that results in animals leaving experimental units may increase abundances on nearby controls. Additionally, if time is the grouping factor and the data are repeated observations on the same experimental or sampling units, then there will often be positive correlations between observations through time.

This assumption must usually be considered at the design stage. Note that we are not arguing that non-independent observations preclude statistical analysis. However, if standard linear models are to be used, it is important to ensure that experimental units are independent of each other, both within and between groups. Randomization at various stages of the design process can help provide independence, but can't guarantee it. For some specific designs, hypothesis tests for linear models can be adjusted conservatively to correct for increasing Type I error rates resulting from positive autocorrelations (Chapters 10 and 11), and these tests fall in the general category of unified mixed linear models (Laird & Ware 1982). Alternatively, the lack of independence (e.g. spatial correlations) can be incorporated into the design and the modeling (Legendre 1993, Ver Hoef & Cressie 1993). Generalized estimating equations (GEEs; see Chapter 13), as developed for handling correlated observations in regression models, may also be useful for ANOVA models because they can be applied to models with both continuous and categorical predictors (Ware & Liang 1996). Finally, measuring and testing hypotheses about spatial patterns, especially when we anticipate that sampling units closer together will be more correlated, are more suited to the fields of spatial statistics and geostatistics (see Koenig 1999, Manly 2001 and Rossi *et al.* 1992 for ecological perspectives).

8.4 | ANOVA diagnostics

The predictor variable (factor) in a single factor ANOVA model is categorical, so the range of available diagnostics to check the fit of the model and warn about influential observations is limited compared with linear regression models (Chapter 5). Leverage ("outliers" in the *X*-variable) has no useful meaning for a categorical predictor and Cook's D_i is also hard to interpret, partly because it is based on leverage. However, residuals are still a crucial part of checking any linear model (Box 8.1 and Box 8.2, Table 8.7). Studentized residuals, the residual divided by its standard deviation (Neter *et al.* 1996), are usually easier for comparison between different models because they have constant variance (Table 8.7). Plots of residuals or studentized residuals against predicted values (group means) are the most informative diagnostic tools for ANOVA models. The residuals should show equal spread across groups, indicating variance homogeneity. Increasing spread of residuals (a wedge shape) suggests a skewed (non-normal) distribution of *Y* in each group and unequal variances (Figure 8.5). These residual plots can also indicate autocorrelation, just as in regression analysis (Chapter 5).

Outliers can either be observations very different from the rest in a sample (Chapter 4) or observations with a large residual that are a long way from the fitted model compared with other observations (Chapter 5). Outliers will usually have undue influence on estimates of group effects (or variances) and the conclusions from the ANOVA *F* test. Such observations should always be checked. If they are not a mistake and cannot be corrected or deleted, then one solution is to fit the ANOVA model twice, with the outlier(s) omitted and with

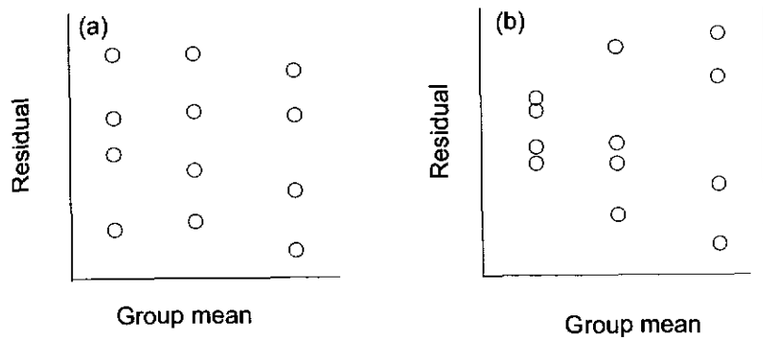

Figure 8.5 Diagrammatic residual plots for single factor ANOVA with three groups and *n* equals four: (a) similar variances, (b) variance increasing with mean, suggesting positively skewed distribution.

the outlier(s) included (Chapter 4). If there is little difference between the two analyses, it suggests that the outlier(s) are not influential. In the worst case, the outlier(s) may change the result between significant and non-significant. There is not much that you can do in this case, other than to describe both results, and discuss biological explanations for the outliers.

8.5 | Robust ANOVA

We have already pointed out that the *F* test in a single factor ANOVA is sensitive to large differences in within group variances, especially when sample sizes are unequal. This has led to the development of a number of alternative tests for differences between groups that are more robust to either heterogeneity of variances or outliers or both. We won't present formulae for these tests because, except for the rank versions, the computations are reasonably tedious and we figure biologists are unlikely to use these tests until they appear in statistical software. These robust tests fall into three categories.

8.5.1 Tests with heterogeneous variances

A number of procedures have been developed for testing equality of group means (and specific comparisons; Section 8.6) when variances are very different (Wilcox 1987a, 1993). One of the earliest was Welch's test (Day & Quinn 1989, Wilcox 1993), which uses adjusted degrees of freedom to protect against increased Type I errors under variance heterogeneity. Wilcox (1997) described a modification of Welch's test that extends Yuen's use of trimmed means to more than two groups (the Yuen–Welch test). Other tests include the Brown–Forsythe test and James second order method (see review by Coombs *et al.* 1996). These tests generally have less power than the standard ANOVA *F* test and are relatively complicated to calculate. Wilcox (1993) described the *Z* test, which is an extension of his *H* test (Chapter 3) for two groups; it is based on *M*-estimators and bootstrap methods to determine critical values. Coombs *et al.* (1996) recommended this test for unequal variances and non-normal distributions and the James second-order method for normal distributions. These tests are not yet available in most statistical software. Our preference is to examine outlying values and, if appropriate, apply a sensible transformation of the data. This encourages researchers to explore their data and think carefully about the scale of measurement. Alternatively, if the underlying distribution of the observations and the residuals is known, and hence why a transformation might be effective, generalized linear modeling (GLM) can be applied (Chapter 13).

8.5.2 Rank-based ("non-parametric") tests

For non-normal distributions (but similar variances), methods based on ranks (Chapter 3) might be used. There are two broad types of rank-based tests for comparing more than two groups. First is the Kruskal–Wallis test, which is a rank-randomization test and an extension of the Mann–Whitney–Wilcoxon test described in Chapter 3 for comparing two groups. It tests the H_0 that there is no difference in the location of the distributions between groups or treatments and is based on ranking the pooled data, determining the rank sums within each group and calculating the *H* statistic that follows a χ^2 distribution with $(p-1)$ df (Hollander & Wolfe 1999, Sokal & Rohlf 1995). Although the Kruskal–Wallis test is non-parametric in the sense that it does not assume

that the underlying distribution is normal, it does assume that the shapes of the distributions are the same in the different groups (the only possible difference being one of location, as tested in the H_0). This implies that variances should be similar (Hollander & Wolfe 1999). Therefore, the Kruskal–Wallis test is not a recommended solution for testing under unequal variances. However, it is a useful approach for dealing with outliers that do not represent more general variance heterogeneity. The Kruskal–Wallis test is sometimes described as a "non-parametric ANOVA" but this is a little misleading; there is no partitioning of variance and the H_0 does not test means unless the distributions are symmetric.

In the rank transform (RT) method, we transform the data to ranks and then fit a parametric ANOVA model to the ranked data (Conover & Iman, 1981). This really is an "analysis of variance" because the RT approach can be viewed as just an extreme form of transformation resulting in an ANOVA on rank-transformed data. It turns out that, for a single factor design, an RT F test will produce the same result as the Kruskal–Wallis test (Neter *et al.* 1996), but it is a more general procedure and can potentially be used for complex ANOVAs (Chapter 9). The RT approach also does not deal with unequal variances; if the variances are unequal on the raw scale, the ranks may also have unequal variances. We would also need to conduct the usual model-checking diagnostics on the ranked data.

8.5.3 Randomization tests

We can also use a randomization test to test the H_0 of no difference between groups (Crowley 1992, Edgington 1995, Manly 1997; Chapter 3). The procedure randomly allocates observations (or even residuals) to groups (keeping the same sample sizes) many times to produce the distribution of a test statistic (e.g. F-ratio or SS_{Groups} or MS_{Groups}; see Manly 1997) under the H_0 of no group effects. If this H_0 is true, we would expect that all randomized allocations of observations to groups are equally likely. We simply compare our observed statistic to the randomized distribution of the statistic to determine the probability of getting our observed statistic, or one more extreme, by chance. Manly (1997) indicated, based on simulations, that randomization of observations and residuals produced similar results. However, such a randomization test to compare group means may not be robust against unequal variances, as Crowley (1992) and Manly (1997) have both pointed out that the H_0 can be rejected because of different variances without any differences between the means. While the conclusions from a randomization test also cannot easily be extrapolated to a population of interest, in contrast to the traditional approaches, randomization tests don't rely on random sampling from populations and therefore will be useful when random sampling is not possible (Ludbrook & Dudley 1998).

8.6 Specific comparisons of means

Very few aspects of applied statistics have created as much discussion and controversy as multiple comparisons, particularly comparisons of group means as part of ANOVAs (see reviews by Day & Quinn 1989, Hancock & Klockars 1996, Hochberg & Tamhane 1987). We discussed the issues related to multiple significance testing in Chapter 3. Much of the debate and development of techniques for dealing with this problem have arisen in the context of multiple comparisons of group means following ANOVA models. Two issues are of particular importance. The first is the general multiple testing problem and increased rate of Type I errors (Chapter 3). For example, if we have five groups in our design, we would need ten pairwise tests to compare all groups with each other. The probability of at least one Type I error among the family of ten tests, if each test is conducted at α equals 0.05 and the comparisons are independent of each other, is 0.40 (Table 3.1). Much of the discussion about specific group contrasts in ANOVA models has focused on the need to correct for this increase in family-wise Type I error rates and the best methods to achieve this correction.

The second issue is independence (the statistical term is orthogonality) of the contrasts. For example, say we have three groups with equal sample sizes and our sample mean for group A is greater than that for group B with group C having the smallest mean. If our pairwise tests reject the null hypotheses that group A and B have equal

population means and that group B and C have equal population means, then the contrast of groups A and C is redundant. We know that the mean for group A is significantly greater than the mean for group C without needing to do the test. Ensuring a set of contrasts is independent (orthogonal) is important for two reasons. First, independent contrasts are straightforward to interpret because the information each contains is independent. Second, the family-wise Type I error rate can be easily calculated if necessary, using Equation 3.9; the family-wise Type I error rate cannot be easily calculated for non-independent contrasts.

As discussed in Chapter 3, traditional adjustments to significance levels to correct for multiple testing are very severe, restricting the family-wise Type I error rate to the same level as the comparison-wise level for each comparison (e.g. 0.05). It seems strange to us that we are willing to allocate a significance level of 0.05 for individual comparisons but as soon as we consider a family of comparisons, we constrain the probability of at least one Type I error to the same level (0.05). The cost of this very tight control of family-wise Type I error is that our individual comparisons have decreasing power as the number of comparisons in our family increases.

Our broad recommendation is that the default position should be no adjustment for multiple testing if the tests represent clearly defined and separate hypotheses (Chapter 3). The exception is when we are scanning all possible comparisons in an exploratory manner where the aim is to detect significant results from all possible tests that could be carried out on a data set. Under these circumstances, we agree with Stewart-Oaten (1995) that some protection against increasing Type I error rates should be considered and, when comparing all possible group means in an ANOVA design, the procedures outlined in Section 8.6.2 should be adopted. However, we also urge biologists not to be constrained to the convention of 0.05 as a significance level. A sensible balance between power and Type I error rate in situations where adjustments are made for multiple testing can also be achieved by setting family-wise Type I error rates at levels above 0.05.

8.6.1 Planned comparisons or contrasts

These are interesting and logical comparisons (often termed contrasts) of groups or combinations of groups, each comparison commonly using a single df. They are usually planned as part of the analysis strategy before the data are examined, i.e. the choice of contrasts is best not determined from inspection of the data (Day & Quinn 1989, Ramsey 1993, Sokal & Rohlf 1996). Many texts recommend that planned contrasts should be independent of each other, where the comparisons should contain independent or uncorrelated information and represent a non-overlapping partitioning of the SS_{Groups}. The number of independent comparisons cannot be more than the df_{groups} $(p-1)$. This means that the outcome of one comparison should not influence the outcome of another (see Maxwell & Delaney 1990 for examples) and the family-wise Type I error rate can be easily calculated (Chapter 3). Even the question of orthogonality is not without differences of opinion among statisticians, and some argue that the set of planned comparisons need not be orthogonal (e.g. Winer *et al.* 1991), and that it is more important to test all of the hypotheses of interest than to be constrained to an orthogonal set. We agree with Toothaker (1993) that orthogonality has been given too much emphasis in discussions of group comparisons with ANOVA models, especially in terms of error rates, and that it is more important to keep the number of contrasts small than worrying about their orthogonality.

There is some consensus in the literature that each planned contrast, especially when they are orthogonal, can be tested at the chosen comparison-wise significance level (e.g. equals 0.05), and no control over familywise Type I error rate is necessary (Day & Quinn 1989, Kirk 1995, Sokal & Rohlf 1995). We agree, although we place much less emphasis on the need for orthogonality. The arguments in favour of not adjusting significance levels are that the number of comparisons is small so the increase in family-wise Type I error rate will also be small and each comparison is of specific interest so power considerations are particularly important. Another argument is that contrasts represent independent hypotheses, so there is no multiple testing involved. This approach is not

universally supported. For example, Ramsey (1993) argued that the family-wise error rate should be controlled in any multiple testing situation, although the power of individual tests in complex ANOVAs with numerous hypotheses (main effects and interactions) surely would be unacceptable with this strategy. Additionally, some statisticians have argued that adjustment is necessary only when non-orthogonal contrasts are included and Keppel (1991) proposed adjusting the significance level of only those comparisons that are not orthogonal – see Todd & Keough (1994) for an example of this approach.

The H_0 being tested is usually of the form $\mu_A = \mu_B$ (e.g. the mean of group A equals the mean of group B). The hypotheses can be more complicated, such as $(\mu_A + \mu_B)/2 = \mu_C$ (e.g. the average of the means of group A and group B equals the mean of group C); note that this comparison will still have one df because there are only two "groups", one being formed from a combination of two others. For example, Newman (1994) examined the effects of changing food levels on size and age at metamorphosis of tadpoles of a desert frog. He used small plastic containers as the experimental units, each with a single tadpole. There were four treatments: low food (one-quarter ration, n equals 5 containers), medium food (half-ration, n equals 8), high food (full ration, n equals 6), and food decreasing from high to low during the experiment (n equals 7). Single factor ANOVAs were used to test for no differences between the four treatments on size and age at metamorphosis. In addition to the overall effect of food level, Newman (1994) was particularly interested in the hypothesis that a deteriorating growth environment changes timing of metamorphosis compared to a constant good environment so he included a single planned contrast: decreasing food vs constant high food.

Another example is from Marshall & Keough (1994), who examined the effects of increasing intraspecific densities of two size classes (large and small) of the intertidal limpet *Cellana tramoserica* on mortality and biomass of large limpets. There were seven treatments (one large limpet per experimental enclosure, two large, three large, four large, one large and ten small, one large and 20 small, one large and 30 small), four replicate enclosures for each treatment and a single factor ANOVA was used to test for treatment effects. They included three specific contrasts to test for the effects of small limpets on large limpets: ten small vs ten small and one large, ten small vs ten small and two large, ten small vs ten small and three large.

In our worked example, Keough & Raimondi (1995) used three specific contrasts to identify effects of different kinds of biofilms on settlement of serpulid worms: lab netted vs lab un-netted, then, with a non-significant result, the average of the two lab films were compared to field biofilms, and, finally, if these did not differ, the average of all three filmed treatments were compared to the substrata that had been in sterile seawater (Box 8.4).

There are two ways of doing these planned comparisons, partitioning the between groups SS or using two group t tests.

Partitioning SS

The SS_{Groups} can be partitioned into the contribution due to each comparison. Each $SS_{Comparison}$ will have one df (we are only comparing two means or two combinations of means) and, therefore, the $SS_{Comparison}$ equals the $MS_{Comparison}$. The H_0 associated with each comparison is tested with an F-ratio ($MS_{Comparison}/MS_{Residual}$). This approach is simply an extension of the partitioning of the variation that formed the basis of the ANOVA and is illustrated in Box 8.4.

We need to define a linear combination of the p means representing the specific contrast of interest:

$$c_1\bar{y}_1 + \ldots + c_i\bar{y}_i + \ldots \text{etc.} \qquad (8.8)$$

where c_i are coefficients, $\sum_{i=1}^{p} n_i c_i$ equals zero (this ensures a valid contrast) and $\bar{y}_i$ are treatment or group means. The details for working out these linear contrasts are provided for the worked example from Keough & Raimondi (1995) in Box 8.4. Note that the absolute values of the coefficients are not relevant as long as the coefficients sum to zero and represent the contrast of interest. We prefer to use integers for simplicity. We can also define orthogonality in terms of coefficients. Two comparisons, A and B, are independent (orthogonal) if $\sum_{i=1}^{p} c_{iA}c_{iB}$ equals zero, i.e. the sum of the products of their coefficients equals zero. It is often not intuitively obvious whether two comparisons are orthogonal, and the only way to be sure is to do these calculations. If comparisons are orthogonal, then the sum of the $SS_{Comparison}$ will not exceed SS_{Groups}. If comparisons are not orthogonal, then sum of $SS_{Comparison}$ can exceed available SS_{Groups}, indicating we are using the same information in more than one comparison.

t test

A t test can be used to compare groups A and B with the modification that

$$\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right) MS_{Residual}}$$

is used as the standard error of the comparison. This standard error makes use of the better estimate of residual variance from the ANOVA (if the assumption of homogeneity of variance holds) and has more df than the usual t test which would just use the data from the two groups being compared.

Partitioning the SS and t test are functionally equivalent; the F-ratio will equal t^2 and the P values will be identical. Both approaches can handle unequal sample sizes and the t test approach can be adjusted for unequal variances (Chapter 3). We prefer the former because it is a natural extension of the ANOVA and the results are easy to present as part of the ANOVA table. Note that a significant ANOVA F test is not necessary before doing planned comparisons. Indeed, the ANOVA might only be done to provide the $MS_{Residual}$.

8.6.2 Unplanned pairwise comparisons

Now we will consider multiple testing situations, and specifically multiple comparisons of means, where control of family-wise Type I error rate might be warranted. There are two broad approaches to adjusting significance levels for multiple testing. The first is to use specific tests, often based on the F or q distributions. A more general method, which can be used for any family tests, is to adjust the P values (Chapter 3).

Unplanned pairwise comparisons, as the name suggests, compare all possible pairs of group means (i.e. each group to every other group) in a *post hoc* exploratory fashion to find out which groups are different after a significant ANOVA F test. These multiple comparisons are clearly not independent (there are more than $p-1$ comparisons), there are usually lots of them and they involve data snooping (searching for significant results, or picking winners (Day & Quinn 1989), from a large collection of tests). There seems to be a much stronger argument that, in these circumstances, some control of family-wise Type I error rate is warranted. The usual recommendation is that the significance level (α) for each test is reduced so the family-wise Type I error rate stays at that chosen (e.g. 0.05).

Underwood (1997) has argued that there has been too much focus on Type I error rates at the expense of power considerations in multiple comparisons. We agree, although controlling family-wise error rates to a known maximum is important. We do not support increasing power of individual comparisons by using procedures that allow a higher, but unknown, rate of Type I errors under some circumstances (e.g. SNK or Duncan's tests – see below). To increase power when doing all pairwise comparisons, we would prefer using a multiple comparison procedure that has a known upper limit to its family-wise error rate and then setting that rate (significance level) above 0.05.

There are many unplanned multiple comparison tests available and these are of two broad types. Simultaneous tests, such as Tukey's test, use the value of the test statistic based on the total number of groups in the analysis, irrespective of how many means are between any two being compared. These simultaneous tests also permit simultaneous confidence intervals on differences between means. Stepwise tests use different values of the test statistic for comparisons of means closer together and are generally more powerful, although their control of the family-wise Type I error rate is not always strict. Both types of test can handle unequal sample sizes, using minor modifications, e.g. harmonic means of sample sizes. Day and Quinn (1989) and Kirk (1995) provide detailed evaluation and formulae but brief comments are included below.

Tukey's HSD test

A simple and reliable multiple comparison is Tukey's (honestly significant differenced, or HSD)

test, which compares each group mean with every other group mean in a pairwise manner and controls the family-wise Type I error rate to no more than the nominal level (e.g. 0.05). Tukey's HSD test is based on the studentized range statistic (q), which is a statistic used for multiple significance testing across a number of means. Its sampling distribution is defined by the number of means in the range being compared (i.e. the number of means between the two being compared after the means are arranged in order of magnitude) and the $df_{Residual}$. The q distribution is programmed into most statistical software and critical values can be found in many textbooks.

We illustrate the logic of Tukey's HSD test as an example of an unplanned multiple comparison test (see also Day & Quinn 1989 and Hays 1994 for clear descriptions):

- As we did for planned comparisons using a t test, calculate the standard error for the difference between two means from

$$\sqrt{\left(\frac{1}{n_A}+\frac{1}{n_B}\right)MS_{Residual}}$$

  Using the harmonic mean of the sample sizes is sometimes called the Tukey–Kramer modification (Day & Quinn 1989) and reduces to $1/n$ for equal sample sizes.
- Determine appropriate value of (q) from the q distribution at the chosen significance level (for family-wise Type I error rate), using $df_{Residual}$ and the number of means being compared (i.e. the number of groups, a).
- Calculate the *HSD* (honestly significant difference, sometimes termed the minimum significant difference, *MSD*; Day & Quinn, 1989). The *HSD* is simply q times the standard error and is the smallest difference between two means that can be declared significant at the chosen family-wise significance level.
- Compare the observed difference between two sample means to the *HSD*. If the observed difference is larger, then we reject the H_0 that the respective population means are equal. Repeat this for all pairs of means.
- Presenting the results of multiple comparisons is not straightforward because the number of tests can be large. Two common approaches are to join those means not significantly different with an underline (e.g. $\underline{A\,B} < \underline{C\,D}$) or to indicate groups not significantly different from each other with the same subscript or superscript in tables or figures.

Fisher's Protected Least Significant Difference test (LSD test)
This test is based on pairwise t tests using pooled within groups variance ($MS_{Residual}$) for the standard error, as described for planned comparisons (Section 8.6.1) and applied only if the original ANOVA F test is significant (hence "protected"). However, it does not control family-wise Type I error rate unless the true pattern among all groups is that there are no differences. It is not recommended for large numbers of unplanned comparisons (Day & Quinn 1989).

Duncan's Multiple Range test
This stepwise test based on the q statistic for comparing all pairs of means was historically popular. However, it does not control the family-wise Type I error rate at a known level (nor was it ever designed to!) and is not recommended for unplanned pairwise comparisons.

Student–Neuman–Keuls (SNK) test
This test is very popular, particularly with ecologists, because of the influence of Underwood's (1981) important review of ANOVA methods. It is a relatively powerful stepwise test based on the q statistic. Like the closely related Duncan's test, it can fail to control the family-wise Type I error rate to a known level under some circumstances when there are more than three means (specifically when there are four or more means and the true pattern is two or more different groups of two or more equal means). Although Underwood (1997) argued that the SNK test might actually be a good compromise between Type I error and per comparison power, we prefer other tests (Tukey's, Ryan's, Peritz's) because they provide *known* control of family-wise Type I error. Underwood (1997) provided formulae and a worked example of the SNK test.

Ryan's test
This is one of the most powerful stepwise multiple comparison procedures that provides control over the family-wise Type I error rate and is often referred to as the Ryan, Einot, Gabriel and Welsch (REGW) procedure. It can be used with either the q or the F-ratio statistic and it is the recommended multiple comparison test if software is available, but it is a little tedious to do by hand.

Peritz's test
This is basically an SNK test that switches to a REGW-type test in situations where the SNK cannot control Type I error, so it is a combined SNK and Ryan's test. It is probably too complicated for routine use.

Scheffe's test
This is a very conservative test, based on the F-ratio statistic, designed for testing comparisons suggested by the data. It is not restricted to pairwise comparisons, in contrast to Tukey's test, but is not very efficient for comparing all pairs of means.

Dunnett's test
This is a modified t test designed specifically for comparing each group to a control group. Under this scenario, there are fewer comparisons than when comparing all pairs of group means, so Dunnett's test is more powerful than other multiple comparison tests in this situation.

Robust pairwise multiple comparisons
Like the ANOVA F test, the multiple comparison tests described above assume normality and, more importantly, homogeneity of variances. Pairwise multiple comparison procedures based on ranks of the observations are available (Day & Quinn 1989) and there are also tests that are robust to unequal variances, including Dunnett's T3, Dunnett's C and Games–Howell tests (Day & Quinn 1989, Kirk 1995). They are best used in conjunction with robust ANOVA methods described in Section 8.5.

Tests based on adjusting *P* values
Multiple comparisons of group means are simply examples of multiple testing and therefore any of the P value adjustment methods described in Chapter 3 can be applied to either t or F tests (or the robust procedures) used to compare specific groups. Sequential Bonferroni methods are particularly appropriate here.

8.6.3 Specific contrasts versus unplanned pairwise comparisons

A small number of planned contrasts is always a better approach than comparing all pairs of means with an unplanned multiple comparison procedure. In most cases, you probably have in mind some specific hypotheses about the groups and tests of these hypotheses are usually more powerful because there is less of an argument for adjusting error rates and they are nearly always more interpretable. They also encourage biologists to think about their hypotheses more carefully at the design stage. Unplanned comparisons are usually only done when the ANOVA F test indicates that there is a significant result to be found; then we often wish to go "data-snooping" to find which groups are different from which others.

As we have already discussed, the adjustment of Type I error rates for most standard unplanned multiple comparison procedures means that the power of individual comparisons can be weak, especially if there are lots of groups, and this can make the unplanned tests difficult to interpret. For example, an unplanned multiple comparison test (with family-wise adjustment), following a "marginally significant" ($0.01 < P < 0.05$) ANOVA F test may not reveal any differences between groups. Also, some unplanned multiple comparisons can produce ambiguous results, e.g. with three means in order smallest (A) to largest (C), the test might show C>A but A=B and B=C! Underwood (1997) has argued that no conclusions can be drawn from such a result because no alternative hypothesis can be unambiguously identified. We view the multiple comparison test as a set of different hypotheses and suggest that such a result allows us to reject the H_0 that A equals C, but no conclusion about whether B is different from A or C.

Finally, specific contrasts of groups are mainly relevant when the factor is fixed, and we are specifically interested in differences between group means. When the factor is random, we are more interested in the added variance component (Section 8.2.1) and not in specific differences between groups.

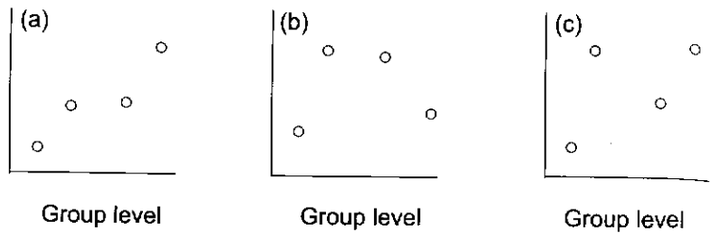

Figure 8.6 Diagrammatic representation of (a) linear, (b) quadratic and (c) cubic trends in *Y* across four equally spaced, quantitative, groups.

8.7 Tests for trends

If the factor in an ANOVA is fixed and quantitative (i.e. the treatment levels have some numerical value), then tests for trends in the group means may be more informative than tests about whether there are specific differences between group means. Usually, we wish to test for a linear trend or some simple nonlinear (e.g. quadratic or cubic – see Chapter 6) trend. For example, Glitzenstein *et al.* (1995) studied the mortality of sandhill oak (*Quercus* spp.) across eight season-of-burn treatments (eight two-week periods during 1981/1982). They used a single factor ANOVA (season-of-burning, with seven df) but were more interested in trends in oak mortality through the eight seasons than specific differences between season means. They tested for linear and quadratic patterns in mortality across burn season. Marshall & Keough (1994) examined the effects of increasing intraspecific densities of two size classes (large and small) of the intertidal limpet *Cellana tramoserica* on mortality and biomass of large limpets. There were seven treatments (one large, two large, three large, four large, one large and ten small, one large and 20 small, one large and 30 small), four replicate enclosures for each treatment and a single factor ANOVA was used to test the overall H_0 of no treatment differences. Marshall & Keough (1994) also included a trend analysis to test for a linear relationship in mean mortality (or biomass) across the intra-size-class treatments (one, two, three, four large limpets per enclosure). We will illustrate a linear trend analysis with the data on serpulid recruitment from Keough & Raimondi (1995), where the equally spaced levels of biofilm represented an increasingly stronger cue for settlement (Box 8.4). Note that the factor is not really quantitative in this example, but can be considered a rank order and therefore still suitable for testing trends.

The method of orthogonal polynomials fits polynomial equations to the group means using contrast coefficients, just as for planned contrasts between specific group means. A linear polynomial represents a straight-line relationship through the group means, a quadratic represents a U-shaped relationship with a single "change of direction" and the cubic represents a more complex pattern with two "changes of direction" (Figure 8.6; Kirk 1995). We don't provide computational details for fitting orthogonal polynomials to group means (see Kirk 1995, Maxwell & Delaney 1990, Winer *et al.* 1991) but the logic is similar to contrasts of means described in Section 8.6.1 and they are simple to do with most statistical software. The SS_{Groups} is partitioned up into SS_{Linear}, $SS_{Quadratic}$, SS_{Cubic}, etc., each with one df. The null hypothesis of no linear (or quadratic, etc.) trend is tested with *F* tests, using the $MS_{Residual}$. Our experience is that polynomials above cubic are difficult to interpret biologically and are rarely fitted in practice, even when there are enough df to do so. If the levels of the factor are equally spaced and sample sizes are equal, the coefficients for the contrasts equations for linear, quadratic, etc., trends can be found in Table 8.8; unequal sample sizes and/or spacing of factor levels are discussed below.

The rules for contrast coefficients still apply. The coefficients for each polynomial should sum to zero and we could multiply the coefficients for any contrast by a constant and still get the same result (e.g. −30, −10, 10, 30 is the same linear contrast as −3, −1, 1, 3). Successive polynomials (linear, quadratic, etc.) are independent (orthogonal) as long as the number of successive polynomials, starting with linear, doesn't exceed the df_{Groups}, i.e. if there are four groups, there are

Table 8.8 Coefficients for linear, quadratic and cubic polynomials for between three and six equally spaced group levels. See Kirk (1995) or Winer *et al.* (1991) for more orders and levels

	X_1	X_2	X_3	X_4	X_5	X_6
Linear	−1	0	1			
	−3	−1	1	3		
	−2	−1	0	1	2	
	−5	−3	−1	1	3	5
Quadratic	1	−2	1			
	1	−1	−1	1		
	2	−1	−2	−1	2	
	5	−1	−4	−4	−1	5
Cubic	−1	3	−3	1		
	−1	2	0	−2	1	
	−5	7	4	−4	−7	5

three df and we can have three orthogonal polynomials: linear, quadratic, cubic.

When sample sizes are equal in all groups, the SS from fitting a linear contrast across the means using orthogonal polynomials will be the same as the $SS_{Regression}$ from fitting a linear regression model to the original observations. Note that the $SS_{Residual}$, and therefore the test of linearity, will be different in the two cases because the classical regression and ANOVA partitions of SS_{Total} are different. The $SS_{Residual}$ from fitting the ANOVA model will be smaller but also have fewer df, as only one df is used for the regression but $(p-1)$ is used for the groups. The difference in the two $SS_{Residual}$ (from the regression model and the ANOVA model) is termed "lack-of-fit" (Neter *et al.* 1996), representing the variation not explained by a linear fit but possibly explained by nonlinear (quadratic, etc.) components.

The SS from fitting a quadratic contrast across the means will be the same as the SS_{Extra} from fitting a quadratic regression model over a linear regression model to the original observations (Chapter 6). So the quadratic polynomial is testing whether there is a quadratic relationship between the response variable and the factor over and above a linear relationship, the cubic polynomial is testing whether there is a cubic relationship over and above a linear or quadratic relationship, and so on. Sometimes, the remaining SS after SS_{Linear} is extracted from SS_{Groups} is used to test for departures from linearity (Kirk 1995).

When sample sizes are unequal or the spacing between factor levels is unequal, contrast coefficients can be determined by solving simultaneous equations (Kirk 1995) and good statistical software will provide these coefficients. Alternatively, we could simply fit a hierarchical series of polynomial regression models, testing the linear model over the intercept-only model, the quadratic model over the linear model, etc. (Chapter 6). Unfortunately, the equality of the SS due to a particular contrast between group means and the SS due to adding that additional polynomial in a regression model fitted to the original observations breaks down when sample sizes are different (Maxwell & Delaney 1990) so the two approaches will produce different (although usually not markedly) results. We prefer using the contrast coefficients and treating the test for a linear trend as a planned contrast between group means.

8.8 Testing equality of group variances

It may sometimes be of more biological interest to test for differences in group variances, rather than group means, when we expect that experimental treatments would affect the variance in our response variable. Tests on group variances may also be a useful component of diagnostic checks of the adequacy of the ANOVA model and the assumption of homogeneity of variance (Section 8.3).

Traditional tests for the H_0 of equal population variances between groups include Bartlett's test, which is based on logarithms of the group variances and uses a χ^2 statistic, Hartley's F_{max} test, which is based on an *F*-ratio of the largest to the smallest variance, and Cochran's test, which is the ratio of the largest variance to the sum of the variances. Unfortunately, Conover *et al.* (1981) and Rivest (1986) have shown that all these tests are very sensitive to non-normality. Given the prevalence of skewness in biological data, this lack of robustness is a serious concern and these tests cannot be recommended for routine use.

Alternative tests recommended by Conover *et al.* (1981) basically calculate new (pseudo)observations that represent changes in the variance and then analyze these pseudo-observations (Ozaydin *et al.* 1999). Levene's test is based on absolute deviations of each observation from its respective group mean or median (i.e. absolute residuals) and is simply an *F* test based on using these absolute deviations in a single factor ANOVA. The H_0 is that the means of the absolute deviations are equal between groups. Although Levene's test is robust to non-normality of the original variable (Conover *et al.* 1981), the pseudo-observations are not necessarily normal nor will their variances be equal (assumptions of the *F* test). Suggested solutions have been to use robust methods for single factor ANOVAs to analyze the pseudo-observations (see Section 8.5), such as ranking them (Conover *et al.* 1981) and then modifying the ranks with score functions (Fligner & Killeen 1976) or even using a randomization test, although we have not seen this recommended.

8.9 | Power of single factor ANOVA

The *F*-ratio statistic, under the H_0 of equal group means, follows a central *F* distribution (see Chapter 1). When the H_0 is false, the *F*-ratio statistic follows a non-central *F* distribution. The exact shape of this distribution depends on df_{Groups}, $df_{Residual}$ and on how different the true population means are under H_A. This difference is summarized by the non-centrality parameter (λ), which is defined as:

$$\lambda = \frac{\sum_{i=1}^{p} \alpha_i^2}{\sigma_\varepsilon^2/n} = \frac{n\sum_{i=1}^{p} \alpha_i^2}{\sigma_\varepsilon^2} \qquad (8.9)$$

To determine the power of a single factor ANOVA, we need to calculate λ (or $\phi = \sqrt{(\lambda/p)}$). This requires us to specify the alternative hypothesis (H_A) and to know (or guess) the residual variation. Remember the general formula relating power and effect size that we used in Chapter 7:

$$\text{Power} \propto \frac{ES\sqrt{n}}{\sigma} \qquad (7.5)$$

The non-centrality parameter λ incorporates the effect size [group effects (α_i) squared] and the within group standard deviation σ. We can then calculate power by referring to power charts (e.g. Neter *et al.* 1996, Kirk 1995), which relate power to λ or ϕ for different df (i.e. *n*). Alternatively, we can use software designed for the purpose. It is important to note, however, that the formal calculations can vary between different texts and software packages (Box 8.5).

These calculations can be used to:

- determine the power of an experiment *post hoc*, usually after a non-significant result,
- determine the minimum detectable effect size for an experiment *post hoc*, and
- calculate sample size required to detect a certain effect size when planning an experiment.

An example of power calculations is included in Box 8.6 and Underwood (1981, 1997) has also provided worked biological examples. These power calculations are straightforward for two groups but become more difficult with more than two groups. When there are only two groups, the effect size is simply related to the difference between the two means. However, when we have more than two groups, the H_A could, for example, have the groups equally spaced or two the same and one different. These different patterns of means will lead to different values of λ, and, hence, power. The difficulty of specifying H_A becomes greater as the number of groups increases, unless we have a very specific H_A that details a particular arrangement of our groups (e.g. a linear trend across groups).

If the number of groups is not too large, one option is to calculate the power for a range of arrangements of treatments. For example, we can calculate the power characteristics for four different arrangements of groups for a given difference in means (between the largest and smallest), such as groups equally spaced, one group different from all others (which are equal), and so on. An example of such power curves are plotted in Figure 8.7 where for a given effect size you can easily see the range of power values. Note that there is little difference for very large or very small differences between the largest and smallest group means. For planning experiments, it may be enough to know the range of power values

Box 8.6 | Worked example: power analysis for serpulid recruitment onto surfaces with different biofilms

Two of the other response variables in the study of recruitment by Keough & Raimondi (1995), the number of spirorbid worms, and bryozoans in the genus *Bugula*, showed no differences between any of the filming treatments, so power becomes an issue. For the spirorbids, the analysis of variance was as follows.

Source	SS	df	MS	*F*	*P*
Biofilms	0.296	3	0.099	1.624	0.210
Residual	1.458	24	0.061		
Total	1.754	27			

The mean for the unfilmed (SL) treatment was 0.273. We can use this information to look at the power of the test for this species.

If we define the effect size as an increase in settlement of 50% over the value for unfilmed surfaces, our ES = 0.137.

First, let's look at the overall power of the test. Suppose that we wanted to detect any effect of biofilms, in which case, the SL treatment would have a value of 0.273, and the other three would be 0.41. The grand mean would be 0.375, giving estimates of α_i of −0.102, 0.034, 0.034, and 0.034. For these values, $\Sigma_{i=}^{p}\ \alpha_i^2 = 0.014$, and, from the table above, our estimate of σ_ε^2 is 0.061. Using Equation 8.9, $\lambda = (7 \times 0.014)/0.061 = 1.614$, and substituting this value into any software that calculates power, using $df_{Groups} = 3$ and $df_{Residual} = 24$, we get power of 0.143. Remember, power is the probability of statistically detecting this effect size if it occurred. This experiment had little chance of detecting an increase in settlement of 50% above the value for unfilmed surfaces.

To see how our specification of H_A affects power, let's look at the power for a pattern that is one of the hardest to detect using an overall *F* test, a gradual trend from largest to smallest mean. Using the example here, the four means would be 0.271, 0.319, 0.364, and 0.410. Then, the power is 0.117. These differences don't seem very large, mainly because the overall power is so low for this group of polychaetes. For comparison, we can look at the data for the bryozoans. Here, the mean for the sterile treatment is 0.820, and the $MS_{Residual}$ is 0.063. For these bryozoans, our general H_A would produce means of 0.82, 1.23, 1.23, and 1.23 for our four treatments. The non-centrality parameter, λ, is 14.01, giving power of 0.84, so we would feel confident that our non-significant result for this group of animals really represents an effect of less than 50%. If we calculate for the general trend case, the four hypothetical means would be 0.82, 0.96, 1.09, and 1.23, λ = 10.38, and power is 0.70, a drop of 15%.

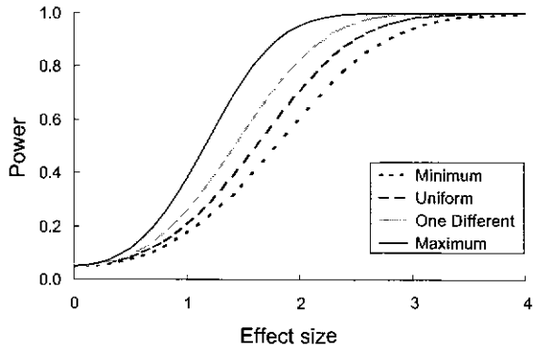

Figure 8.7 Power envelope, showing, for a given effect size, the power for different arrangement of groups. The example used five groups, *n* equals five in each group, standard deviation equals one in each group. Effect size is measured as the difference between the largest and smallest mean.

for a given effect, and to make decisions around one particular arrangement of groups, taking into account where that arrangement fits on the power spectrum.

8.10 General issues and hints for analysis

8.10.1 General issues

- General experimental design principles, especially randomization and choice of appropriate controls, are nearly always important when designing studies for the application of single factor ANOVA models.
- Estimates of explained variance have different interpretations for fixed and random factors. Added variance component for a random factor is straightforward with equal sample sizes and confidence intervals should be used. Explained variance for a fixed factor is also useful as a descriptor but cannot be easily compared for different models and data sets and must be interpreted carefully.
- Be aware that some alternatives that may be suggested as an option when ANOVA assumptions are violated are rarely assumption free. For example, the rank-based non-parametric methods don't assume normality, but have an assumption equivalent to the homogeneity of variances.
- We recommend planned comparisons (contrasts) rather than unplanned multiple comparisons. In most cases, you are not interested in comparing all possible groups, but can identify particular questions that are of greater interest.
- Power calculations are relatively simple for single factor models. However, once the number of groups is greater than two, you must think hard about the kind of differences between groups that is of interest to you. Different alternative patterns of means have different power characteristics.
- A problem for inexperienced biologists is that many of the decisions (how normal should the data be?, etc.) involve an informed judgment about where a particular data set fits along a continuum from assumptions being satisfied completely to major violations. There is no unambiguous division, but, in many cases, it doesn't matter because the *P* values will be far from any grey zone.

8.10.2 Hints for analysis

- Aim for equal sample sizes. The linear model calculations can easily handle unequal samples, but the analysis is more sensitive to the underlying assumptions and parameter estimates and hypothesis tests will be more reliable if sample sizes are equal.
- Homogeneity of variances is an important assumption. ANOVA is robust to small and moderate violations (especially with equal sample sizes), but big differences (e.g. many-fold differences between largest and smallest variances) will alter the Type I error rate of the *F* test.
- Examine homogeneity of variances with exploratory graphical methods, e.g. look at the spread of boxplots, plot group variances or standard deviations against group means, or plot residuals against group means and look for patterns. We don't recommend formal tests of equal group variances as a preliminary check before an ANOVA.
- Transformations will be effective when the error terms, and the observations, have positively skewed distributions. For biological data, the most likely effective transformations are log and square (or fourth) root. Although ANOVA models are robust to violations of non-normality, such normalizing transformations will usually make variances more similar between groups.
- For moderate violations of normality and homogeneity of variances, we recommend proceeding with the analysis, but being cautious about results that are marginally significant or non-significant. Otherwise we recommend using generalized linear models when the underlying distribution of the response variable can be determined, or one of the robust tests.
- Use planned contrasts wherever possible for testing specific differences between groups. If unplanned comparisons must be used, Ryan's (REGW) or Tukey's tests are recommended, the latter if simultaneous confidence intervals are required.