Chapter 9

Multifactor analysis of variance

In Chapter 8, we examined designs with a single factor where the appropriate linear model had a single categorical predictor variable. Commonly in biology, however, we design studies with more than one factor and there are two main reasons why we might include additional factors. First, to try and reduce the unexplained (or residual) variation in our response variable, similarly to multiple regression (Chapter 6). Second, to examine the interactions between factors, i.e. whether the effect of a particular factor on the response variable is dependent on another factor. In this chapter, we will examine two types of multifactor design, nested and factorial, and describe the appropriate linear models for their analysis. The emphasis is on completely randomized (CR) designs, following from Chapter 8, where the experimental units are randomly allocated to factor groups or combinations of factor groups.

9.1 | Nested (hierarchical) designs

A common extension of the single factor design, and the single factor ANOVA linear model, is when additional factors are included that are nested within the main factor of interest. An example based on a manipulative experiment comes from Quinn & Keough (1993) who examined the effect of different enclosure (fence) sizes on growth of the rocky intertidal limpet *Cellana tramoserica*. Part of that experiment used two enclosure sizes (1225 cm^2 and 4900 cm^2), with five replicate enclosures nested within each size and four or five replicate limpets from each enclosure. The response variable was limpet shell height. These nested designs can also be part of sampling programs. For example, Caselle & Warner (1996) looked at recruitment densities of a coral reef fish at five sites on the north shore of the US Virgin Islands, with six random transects within each site and replicate observations of density of recruits along each transect.

Both these examples are two factor nested (or hierarchical) designs, where the levels (categories) of the nested factor are different within each level of the main factor. Quinn & Keough (1993) used enclosure size as the main factor, replicate enclosures within enclosure size as the nested factor and replicate limpets from each enclosure as the residual. Caselle & Warner (1996) used sites as the main factor, transects within each site as the nested factor and replicate observations of fish density as the residual.

The characteristic feature of nested designs that distinguish them from other multifactor designs is that the categories of the nested factor(s) within each level of the main factor are different. The main factor can be fixed or random whereas the nested factor(s) is(are) usually random in biology, often representing levels of subsampling or replication in a spatial or temporal hierarchy. In the example from Quinn & Keough (1993), the enclosures are replicates for the enclosure size treatments, the individual limpets are replicates for the enclosures. However, fixed nested factors can also occur. Bellgrove *et al.* (1997), studying the abundance of algal propagules along exposed rocky coastlines, collected volumes of water from an intertidal shore at different dates within two seasons. The dates within each season were chosen specifically to correspond to the start and end of other experiments (i.e. they were not randomly chosen and so represent a fixed factor) but they were clearly different dates in each of the two seasons (so date was a nested, fixed, factor). Caselle & Warner (1996) also analyzed temporal variation in recruitment of reef fish and chose specific (fixed) months (from the time of the year when the fish recruited) nested within each of two years.

Grazing by sea urchins

To illustrate the formal analysis of nested designs, we will use a recent example from the marine ecological literature. Andrew & Underwood (1993) studied the effects of sea urchin grazing on a shallow subtidal reef in New South Wales, Australia. They set up four urchin density treatments (0% original, 33% original, 66% original, 100% original), with four patches (3–4 m^2) of reef for each treatment and five quadrats from each patch. The response variable was percentage cover of filamentous algae in each quadrat. The complete analysis of these data is in Box 9.1.

Box 9.1 | Worked example of nested ANOVA: grazing by sea urchins

Andrew & Underwood (1993) manipulated the density of sea urchins in the shallow subtidal region of a site near Sydney to test the effects on the percentage cover of filamentous algae. There were four urchin treatments (no urchins, 33% of original density, 66% of original density and 100% of orginal density). The treatments were replicated in four distinct patches (3–4 m^2) per treatment and percentage cover of filamentous algae (response variable) was measured in five random quadrats per patch. This is a nested design with treatment (fixed factor), patch nested within treatment (random factor) and quadrats as the residual.

Null hypotheses

No difference in the mean amount of filamentous algae between the four sea urchin density treatments.

No difference in the mean amount of filamentous algae between all possible patches in any of the treatments.

ANOVA

There were large differences in within-cell variances. Even the variances among patch means within treatments varied, with very low variance among control patch means. These data are percentages, although an arcsin$\sqrt{}$ had no effect in improving variance homogeneity, nor did a log transformation. Like Andrew & Underwood (1993), we analyzed untransformed data, relying on the robustness of tests in balanced ANOVA designs.

Source of variation	df	MS	*F*	*P*	Var. comp.
Treatment	3	4809.71	2.72	0.091	(151.98)
Patches (treatment)	12	1770.16	5.93	<0.001	294.31
Residual	64	298.60			298.60

There was significant variation between the replicate patches within each treatment but there was no significant difference in amount of filamentous algae

between treatments. The very low variances between observations within control patches and between control patch means would cause us concern about the reliability of this analysis. However, when the control group is omitted, the test comparing treatments results in a P value of 0.397. Any treatment effect might be due to the low means of, or the low variance between, control patches compared to the rest, although this analysis cannot separate effects on means from effects on variances. A robust Welch test comparing the four treatment groups, based on patch means, also did not find any significant differences.

The variance in algal cover due to patches was very similar to that due to quadrats within patches. Because the design was balanced, ANOVA, ML and REML all gave identical estimates of components of variance for the random nested factor and the residual. If we equate the mean squares to their expected values and calculate the "variance" component for the fixed treatment effects, we can see that less of the total variation in algal cover was explained by the fixed density effects than by the random patch and quadrat terms.

A one factor ANOVA comparing the four treatments with patch means as replicates produces an identical F test for the main effect (note that the MS values are smaller, by a factor of five, the number of quadrats, but the F-ratios are identical).

Source of variation	df	MS	F	P
Treatment	3	961.9	2.72	0.091
Residual	12	354.0		

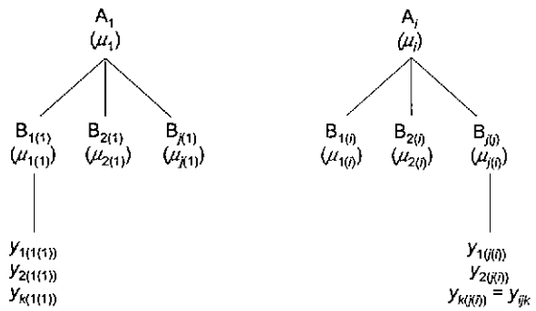

Figure 9.1 Part of data set for two factor nested ANOVA, with p levels of factor A ($i = 1$ to p), q levels of factor B ($j = 1$ to q), where the levels of B are different within each level of A, and n replicate observations within each combination (cell) of A and B nested within A ($k = 1$ to n).

9.1.1 Linear models for nested analyses

Linear effects model

Complex designs can be represented with factor relationship diagrams (Bergerud 1996). Let us consider the two factor nested design, shown in Figure 9.1 and illustrated with the specific example from Andrew & Underwood (1993) in Table 9.1. The main factor A (sea urchin density treatment) has p equals four groups ($i = 1$ to p), the nested factor B (patch) has q equals four groups within each level of A ($j = 1$ to q) and there are n equals five replicate quadrats ($k = 1$ to n) within each combination of A and B categories (patch and density treatment). Note that the groups (levels) of factor B, the patches, are different within each level of A (sea urchin density), so any patch within 0% original density cannot be the same as any patch within 33% original density and so on. Clearly, the same applies to replicate quadrats that are different within each combination of density and patch. Analysis of designs with unequal numbers of levels of B within each level of A, and of replicate observations within each level of B will be discussed in Section 9.1.4.

The mean for each level of A is μ_i (the average of the means for all possible levels of B within each level of A) and the mean for each level of B within each level of A is $\mu_{j(i)}$ (Table 9.1). Note the subscripting, where $j(i)$ represents the jth level of factor B within the ith level of factor A.

Table 9.1 Data structure and sample means for percentage cover of algae from Andrew & Underwood (1993). Factor A was four densities of sea urchins (100%, 66%, 33% and 0% of natural density), factor B was four patches of reef nested within each density treatment and there were n equals five replicate quadrats within each patch within each density

Factor A (A_i)	Density	Density mean $\bar{y}_i$ est μ_i	Factor B ($B_{j(i)}$)	Patch	Patch mean $\bar{y}_{j(i)}$ est $\mu_{j(i)}$
A_1	0%	39.2	$B_{1(1)}$	1	34.2
			$B_{2(1)}$	2	62.0
			$B_{3(1)}$	3	2.2
			$B_{4(1)}$	4	58.4
A_2	33%	19.0	$B_{1(2)}$	5	2.6
			$B_{2(2)}$	6	0.0
			$B_{3(2)}$	7	37.6
			$B_{4(2)}$	8	35.8
A_3	66%	21.6	$B_{1(3)}$	9	28.4
			$B_{2(3}$	10	36.8
			$B_{3(3)}$	11	1.0
			$B_{4(3}$	12	20.0
A_4	100%	1.3	$B_{1(4)}$	13	1.6
			$B_{2(4)}$	14	0.0
			$B_{3(4)}$	15	1.0
			$B_{4(4)}$	16	2.6

The linear (effects) model used to analyze this nested design is:

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk} \quad (9.1)$$

The details of the nested linear ANOVA model, including estimation of its parameters and means, are provided in Box 9.2 and Table 9.2. OLS means and their standard errors are standard output from linear models routines in statistical software and can handle unequal sample sizes.

The model used by Andrew & Underwood (1993) was:

$$(\%\text{ cover algae})_{ijk} = \mu + (\text{sea urchin density})_i + (\text{patch within sea urchin density})_{j(i)} + \varepsilon_{ijk} \quad (9.2)$$

In models 9.1 and 9.2 we have the following.

y_{ijk} is the percentage cover of algae in the kth replicate quadrat from the jth patch within the ith density.

μ is the (constant) mean percentage cover of algae over all possible quadrats in all possible patches in the four sea urchin density treatments.

In this study, sea urchin density is a fixed factor, so α_i is the effect of the ith density, which is the difference between the mean algal cover for the ith sea urchin density treatment and the overall mean algal cover for all the sea urchin density treatments.

Factor B is nearly always random in biology so $\beta_{j(i)}$ is a random variable with a mean of zero and a variance of σ_β^2, measuring the variance among all patches that could have been chosen within each of the four sea urchin density treatments.

ε_{ijk} is residual or unexplained error associated with the kth quadrat within the jth patch within the ith density. This term measures the error associated with each replicate observation (quadrat) of algal cover within each patch within each sea urchin density treatment. The variance of these error terms is σ_ε^2.

The model used by Caselle & Warner (1996) was:

$$(\text{recruit densities})_{ijk} = \mu + (\text{site})_i + (\text{transect within site})_{j(i)} + \varepsilon_{ijk} \quad (9.3)$$

Table 9.2 | OLS estimates of cell and marginal means, with standard errors, in a two factor linear model with equal sample sizes per cell

	Population mean	Sample mean	Standard error
Cell mean	μ_{ij}	$\frac{\sum_{k=1}^{n} y_{ijk}}{n}$	$\sqrt{\frac{MS_{Residual}}{n}}$
Nested design			
Factor A mean	μ_i	$\frac{\sum_{j=1}^{q} \bar{y}_{j(i)}}{q}$	$\sqrt{\frac{MS_{B(A)}}{qn}}$
Crossed design			
Factor A mean	μ_i	$\frac{\sum_{j=}^{q} \bar{y}_{i}}{q}$	$\sqrt{\frac{MS_{Residual}}{qn}}$
Factor B mean	μ_j	$\frac{\sum_{i=1}^{p} \bar{y}_{j}}{p}$	$\sqrt{\frac{MS_{Residual}}{pn}}$

Box 9.2 | The nested ANOVA model and its parameters

The main factor A has p groups ($i = 1$ to p), the nested factor B has q groups within each level of A ($j = 1$ to q) and there are n_{ij} replicates ($k = 1$ to n_{ij}) within each combination of A and B categories. Assume the number of levels of B in each level of A is the same and the number of replicates (n) in each combination of A and B is the same. There are a total of pq cells in this nested design with n replicate observations in each cell. The mean for each level of A is μ_i (the average of the means for all possible levels of B within each level of A) and the mean for each level of B within each level of A is $\mu_{j(i)}$. Note the subscripting, where $j(i)$ represents the jth level of factor B within the ith level of factor A. The linear (effects) model used to analyze this nested design is:

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk} \quad (9.1)$$

In model 9.1 we have the following:

y_{ijk} is the kth replicate observation from the jth group of factor B within the ith group of factor A.

μ is the overall (constant) mean of the response variable.

If factor A is fixed, α_i is the effect of the ith group which is the difference between each A group mean and the overall mean $\mu_i - \mu$. If factor A 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 factor A that could have been used. Factor B is nearly always random in biology so $\beta_{j(i)}$ is a random variable with a mean of zero and a variance of σ_β^2, measuring the variance in mean values of the response variable across all the possible levels of factor B that could have been used within each level of factor A.

ε_{ijk} is residual or unexplained error associated with the kth replicate within the jth level of B within the ith level of A. These error terms are assumed to be normally distributed at each combination of A and B, with a mean of zero ($E(\varepsilon_{ij}) = 0$) and a variance of σ_ε^2.

Model 9.1 is overparameterized (see Box 8.1) because the number of cell means is less than the number of model parameters to be estimated ($\mu, \alpha_1 \ldots \alpha_p, \beta_{1(1)} \ldots \beta_{q(p)}$). In the usual situation of factor A being fixed and factor B being random, estimation of the parameters of the effects model 9.1 can still be achieved using the sum-to-zero constraint $\sum_{i=1}^{p} \alpha_i = 0$, as outlined in Box 8.1. Alternatively, a simpler means model could be fitted:

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

where μ_{ij} is the mean of the response variable for each combination of A and B (each cell). Cell means models don't offer many advantages for nested designs but do become important when we consider missing cells designs in Section 9.2.6.

OLS estimates of the parameters of the nested linear model 9.1 follow the procedures outlined for a single factor model in Chapter 8 with the added complication of two of more factor effects. When the nested factors are random, the means of levels of factor A are estimated from the average of the cell means in each level of A. With different sample sizes within each cell, this results in unweighted means for factor A groups (Table 9.2). OLS standard errors of means in nested designs are calculated using the mean square in the denominator of F-ratio statistic used for testing the H_0 that the means are equal. With A fixed and B(A) random, then $MS_{B(A)}$ will be used for standard errors for factor A means (Table 9.2).

The estimate of the effect of any level of factor A ($\alpha_i = \mu_i - \mu$) is simply the difference between the sample marginal mean for that group and the overall mean:

$$\bar{y}_i - \bar{y}$$

Factor B is usually random, so $\beta_{j(i)}$ is a random variable with a mean of 0 and a variance of σ_β^2 and it is this variance which is of interest, the variance in mean values of the response variable between all the possible levels of factor B that could have been used within each level of factor A. This is estimated as a variance component (Section 9.1.6).

Imagine that Cassele & Warner (1996) had chosen sites at random from a population of possible sites on the north shore of the US Virgin Islands. Then factor A is random and α_i has a mean of zero and a variance of σ_α^2, measuring the variance in the mean number of fish recruits per transect across all the possible sites that could have used in their study.

Predicted values and residuals

The predicted or fitted values of the response variable from model (9.1) are:

$$\hat{y}_{ijk} = \bar{y} + (\bar{y}_i - \bar{y}) + (\bar{y}_{j(i)} - \bar{y}_i) = \bar{y}_{j(i)} \quad (9.4)$$

Any predicted Y-value is predicted by the sample mean for the cell (level of B within each level of A) that contains the Y-value. For example, the predicted percentage cover of algae for quadrat one in patch one for the zero density treatment is the sample mean for patch one for the zero density treatment.

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

$$e_{ijk} = y_{ijk} - \bar{y}_{j(i)} \quad (9.5)$$

Table 9.3 | ANOVA table for two factor nested linear model with factor A (p levels), factor B (q levels) nested within A, and n replicates within each combination of A and B

Source	SS	df	MS
A	$nq\sum_{i=1}^{p}(\bar{y}_i-\bar{y})^2$	$p-1$	$\frac{SS_A}{p-1}$
B(A)	$n\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{y}_{j(i)}-\bar{y}_i)^2$	$p(q-1)$	$\frac{SS_{B(A)}}{p(q-1)}$
Residual	$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{j(i)})^2$	$pq(n-1)$	$\frac{SS_{Residual}}{pq(n-1)}$
Total	$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(y_{ijk}-\bar{y})^2$	$pqn-1$	

For example, the residuals from the model relating algal cover to sea urchin density and patch nested within density are the differences between the observed algal cover on each quadrat and the mean algal cover for the patch and density combination (cell) that contained that quadrat. Note that the sum of the residuals within each cell ($\sum_{k=1}^{n}e_{ijk}$) equals zero. As for all linear models, 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 9.1.7). The OLS estimate of σ_ε^2 is the sample variance of these residuals and is termed the Residual (or Error) Mean Square and is determined as part of the partitioning of the total variation in the response variable described in the next section.

9.1.2 Analysis of variance

The partitioning of the variation in the response variable Y proceeds in a similar manner to that for a single factor model described in Chapter 8. The SS_{Total} in Y can be partitioned into its additive components as illustrated for balanced designs in Table 9.3. These formulae are not really used in practice (and are for balanced designs only), as we estimate the ANOVA terms and test relevant hypotheses by comparing the fit of general linear models (Section 9.1.5). Nonetheless, the formulae in Table 9.3 illustrate the logic behind the partitioning of the total variation in Y.

SS_A measures the sum of squared differences between each A mean and the overall mean, e.g. sum of squared differences between the mean percentage cover of algae for each density treatment and the overall mean percentage cover of algae.

$SS_{B(A)}$ measures the sum of squared differences between each B mean (i.e. cell mean) and the mean of the appropriate level of A, summed across the levels of A, e.g. the sum of squared differences between the mean percentage cover of algae for each patch and the mean percentage cover of algae for the density treatment containing that patch, summed over all density treatments.

$SS_{Residual}$ measures the sum of squared differences between each replicate observation and the appropriate B mean within each cell, summed across all cells, e.g. the sum of squared differences between the percentage cover of algae in each quadrat and the mean percentage cover of algae for the patch containing that quadrat, summed over all patches in all density treatments.

These SS are divided by the appropriate df to produce mean squares (MS or variances). The df_A is simply the number of A levels minus one $[p-1]$, the df_B is the number of B levels within each A level minus one summed over the A levels $[p(q-1)]$ and the $df_{Residual}$ is the number of observations in each cell minus one summed over all cells $[pq(n-1)]$.

Statisticians have determined what population values these sample mean squares estimate,

Table 9.4 | Expected mean squares and F-ratios for tests of null hypotheses for two factor nested ANOVA model

	A fixed, B random		A fixed, B fixed	
Source	Expected mean square	F-ratio	Expected mean square	F-ratio
A	$\sigma_\varepsilon^2+n\sigma_\beta^2+nq\frac{\sum_{i=1}^{p}\alpha_i^2}{p-1}$	$\frac{MS_A}{MS_{B(A)}}$	$\sigma_\varepsilon^2+nq\frac{\sum_{i=1}^{p}\alpha_i^2}{p-1}$	$\frac{MS_A}{MS_{Residual}}$
B(A)	$\sigma_\varepsilon^2+n\sigma_\beta^2$	$\frac{MS_{B(A)}}{MS_{Residual}}$	$\sigma_\varepsilon^2+n\frac{\sum_{i=1}^{p}\sum_{j=1}^{q}\beta_{j(i)}^2}{p(q-1)}$	$\frac{MS_{B(A)}}{MS_{Residual}}$
Residual	σ_ε^2	σ_ε^2		

i.e. what their expected values are, if the assumption of homogeneity of variance (see Section 9.1.7) holds (Table 9.4). In the usual situation of factor A being fixed and factor B random, the $MS_{Residual}$ estimates σ_ε^2 (the variance in the error terms in each cell, pooled across cells), $MS_{B(A)}$ estimates σ_ε^2 plus added variance due to the effects of factor B and MS_A estimates the sum of both these components plus the added effect of fixed levels of factor A.

9.1.3 Null hypotheses

There are two null hypotheses that we test in a two factor nested model, the test for no effects of A and the test for no effects of B nested within A. The expected values of the MS (Table 9.4) provide the logic for testing these null hypotheses, analogous to the single factor model (Chapter 8).

Factor A

$H_0(A)$: $\mu_1=\mu_2=\ldots=\mu_i=\mu$, i.e. no difference between the means for factor A. This is equivalent to $H_0(A)$: $\alpha_1=\alpha_2=\ldots=\alpha_i=0$, i.e. no effect of any level of factor A. In the Andrew & Underwood (1993) example, this null hypothesis is that there is no difference in the mean percentage algal cover between urchin densities. This H_0 is essentially that for a single factor model, using the means for each patch (B level) as replicate observations for the test of urchin density (A level).

If A is random, then $H_0(A)$ is σ_α^2 equals zero, i.e. no added variance due to differences between all the possible levels of A.

Factor B

$H_0(B)$: σ_β^2 equals zero if factor B is random, i.e. no added variance due to differences between all the possible levels of B with any level of A. In the Andrew & Underwood (1993) example, this H_0 is that there is no added variation due to differences in mean percentage algal cover between patches within any urchin density treatment.

In the rarer case of B being a fixed factor, then $H_0(B)$ is $\mu_{1(1)}=\mu_{2(1)}=\ldots=\mu_{j(i)}=\ldots=\mu$, i.e. no difference between the means of the specifically chosen levels of B within any level of factor A. This H_0 when B is fixed is equivalent to H_0: $\beta_{1(1)}=\beta_{2(1)}=\ldots=\beta_{j(i)}=0$, i.e. no effect of any of the specifically chosen levels of factor B within any level of factor A. This is a pooled test of differences between the levels of B for each level of A and the H_0 is false if the mean values for the levels of B are different from each other within one or more of the levels of A.

F-ratios

The F-ratios for testing these H_0s are provided in Table 9.4. If $H_0(A)$ that there is no effect of factor A is true, then all α_is equal zero and MS_A and $MS_{B(A)}$ both estimate $\sigma_\varepsilon^2+n\sigma_\beta^2$ so their ratio (F-ratio) should be less than or equal to one. If $H_0(B)$ that there is no added variance due to differences between the possible levels of factor B within each level A is true, then all β_j equal zero (and therefore $n\sigma_\beta^2$ equals zero) and $MS_{B(A)}$ and $MS_{Residual}$ both estimate σ_ε^2 so their ratio (F-ratio) should be one.

These F-ratios follow an F distribution under

homogeneity of variance and normality assumptions (Section 9.1.7) with one exception. If B is random and the number of replicate observations within each level of B varies, then the F-ratio of $MS_{B(A)}$ and $MS_{Residual}$ does not follow an F distribution when σ_β^2 is greater than zero because $MS_{B(A)}$ is not distributed as a multiple of a χ^2 distribution (Searle *et al.* 1992, see Chapter 8). This also affects estimates of variance components (Chapter 8) and power calculations (Section 9.1.10) for unbalanced nested models. Fortunately, the F-ratio of $MS_{B(A)}$ and $MS_{Residual}$ does follow an F distribution when σ_β^2 equals zero, so the F-ratio test of the H_0 for B(A) with unbalanced data is unaffected.

When B is a random factor, $MS_{B(A)}$ provides the denominator for the F-ratio for the test of A, i.e. the units of replication for testing the effects of A are the means of B. This has important considerations for the power of the test for factor A (Section 9.1.10) and the design of experiments based on nested models. When B is fixed, the expected MS for A does not include a component for B so the F-ratio for testing A uses $MS_{Residual}$ as the denominator. If A is random, the F-ratios are the same as if A is fixed. Note that some statistical software assumes all factors are fixed so will not, by default, provide the correct F tests for nested ANOVAs when the nested factors are random. This problem was pointed out by Ouborg & van Groenendael (1996), who correctly criticized the paper of Heschel & Paige (1995) for incorrectly using the $MS_{Residual}$ instead of $MS_{B(A)}$ in their nested ANOVAs comparing populations of the scarlet gilia (a species of plant), with random seed families nested within populations, and replicates within each seed family (see also response by Paige & Heschel 1996).

9.1.4 Unequal sample sizes (unbalanced designs)

Unequal sample sizes can occur in nested designs in two ways. First, there can be unequal numbers of observations within each cell (unequal n_{ij}). Second, there can be unequal numbers of levels of the nested factor(s) within each level of the higher factor. Neither case is different to unequal sample sizes for single factor ANOVA models and neither causes any computational difficulties. However, as for all linear models fitted by OLS, tests of hypotheses using F-ratios are more sensitive to violations of the assumptions (normality, homogeneity of variances) when sample sizes are unequal (see Chapter 8). Additionally, estimation of variance components for random nested factors is difficult with unequal sample sizes (Chapter 8). When the test for factor A is based on different numbers of B means within each A level, the analysis could be based on a missing cells design and the cell means model used (Kirk 1995; see also Chapter 8). However, as there are no interactions involved, this seems an unnecessary complication.

9.1.5 Comparing ANOVA models

The relative importance of different terms in the linear model for a nested design can be measured, and tests of hypotheses about these terms can also be done, by comparing full and reduced models as described in Section 8.1.5. For example, to test the H_0 that σ_β^2 equals zero, we would compare the fit of the full model (9.1) to a reduced model that omits the B(A) term:

$$y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk} \qquad (9.6)$$

Using the example from Andrew & Underwood (1993), we would compare the fit of model 9.2 to the reduced model:

$$(\%\ \text{cover algae})_{ijk} = \mu + (\text{sea urchin density})_i + \varepsilon_{ijk} \qquad (9.7)$$

The difference in fit of these two models is simply the difference in their $SS_{Residual}$. This difference can be converted to a mean square by dividing by the difference in the $df_{Residual}$. The H_0 of no difference in fit of the two models (i.e. σ_β^2 equals zero; no added variance due to all the possible levels of factor B within each level of factor A) can be tested with an F test using $MS_{Residual}$ of the full model as the denominator. This is, of course, the identical test to that carried out as part of the nested ANOVA.

9.1.6 Factor effects in nested models

The estimation of the effect of the main fixed factor in these nested models is described in Box 9.2, although biologists usually examine fixed factors with planned contrasts or unplanned pairwise comparisons. The estimation of components of variance for random factors in nested models follows the procedures outlined in Chapter 8 for single factor models. The sample mean squares are equated to their expected values (the ANOVA approach) and the added variance due to the nested factors and the residual can be estimated (Table 9.5). The individual variance components for nested models with two or more nested factors are straightforward extensions of those for two factor models once the expected mean squares are known (Table 9.6). Note that these estimates of variance components for random nested factors are only valid for equal sample sizes within each level of the random factor. If the design is unbalanced, estimation of variance components and derivation of confidence intervals is more difficult (Searle *et al.* 1992), although Burdick & Graybill (1992) provide formulae. In general, the REML approach dicussed in Section 8.2 is considered more reliable than the ANOVA method for estimating variance components of random factors above the residual (Searle *et al.* 1992).

Table 9.5 | Estimates of variance components (using ANOVA approach) for two factor nested design with B(A) random

Source	Estimated variance component
A	$\dfrac{MS_A - MS_{B(A)}}{nq}$*
B(A)	$\dfrac{MS_{B(A)} - MS_{Residual}}{n}$
Residual	$MS_{Residual}$

Note:
*This represents variance between population means of specific levels of A if factor A is fixed and a true added variance component if A is random.

Table 9.6 | (a) Estimates of variance components (using ANOVA approach) for three factor nested design with factors A (p levels), B within A (q levels) and C within B within A (r levels) random and n replicates within each cell. (b) Illustration of variance components for nested design from Downes *et al.* (1993) – see Section 9.1.6 for details

(a)

Source	Expected mean square	Estimated variance component	F-ratio
A	$\sigma_\varepsilon^2 + n\sigma_\gamma^2 + nr\sigma_\beta^2 + nrq\sigma_\alpha^2$	$\dfrac{MS_A - MS_{B(A)}}{nrq}$	$\dfrac{MS_A}{MS_{B(A)}}$
B(A)	$\sigma_\varepsilon^2 + n\sigma_\gamma^2 + nr\sigma_\beta^2$	$\dfrac{MS_{B(A)} - MS_{C(B(A))}}{nr}$	$\dfrac{MS_{B(A)}}{MS_{C(B(A))}}$
C(B(A))	$\sigma_\varepsilon^2 + n\sigma_\gamma^2$	$\dfrac{MS_{C(B(A))} - MS_{Residual}}{n}$	$\dfrac{MS_{C(B(A))}}{MS_{Residual}}$
Residual	σ_ε^2	$MS_{Residual}$	

(b)

Source	df	MS	Estimated variance component	% of total variance
Site	2	36188.34	−691.94*	0
Riffle	3	56946.62	2991.15	28
Group	24	12074.12	2103.87	19
Stone (residual)	60	5762.52	5762.52	53

Note:
*Negative variance component converted to zero.

Confidence intervals for σ_ε^2 are calculated in the same way as for single factor designs and work for both balanced and unbalanced designs (Table 8.5). Confidence intervals on the remaining variance components can also be calculated, with approximations based on unweighted SS for unbalanced designs, although the formulae are somewhat tedious (Burdick & Graybill 1992). Note that for a nested model with A fixed and B random, the test of the H_0 that σ_β^2 equals zero is still reliable; it is only when σ_β^2 is greater than zero that the *F*-ratio of $MS_{B(A)}/MS_{Residual}$ no longer follows an *F* distribution, and estimation of a non-zero variance component is difficult.

These nested designs are commonly used to partition the variation in a response variable among levels of a spatial or temporal hierarchy and we are often interested in calculating the relative contribution of random nested terms to the total variation in *Y*. For example, Downes *et al.* (1993) examined spatial variation in the distribution of invertebrates living on stones in a stream. They used three randomly chosen sites (covering about 1.5 km of stream), two riffles (shallow, fast-flowing, stony areas) at each site, five groups of stones from each riffle and three stones from each group and wished to test the relative contribution of each of the spatial scales to the variation in total density of invertebrates. The components of variance for each of the random factors can be estimated using an appropriate method (ANOVA for balanced designs, REML or ML for unblanced designs) and the percentage contribution of each random term to the total variance of the random terms can be calculated (Table 9.5).

In the common situation of a fixed main factor with one or more random nested factors, we can also partition the total variance using the ANOVA approach for both the fixed and nested random factors (Table 9.5). It is very important to remember that the interpretation of the true variance components for B(A) and Residual is quite different from the variance between fixed treatment effects for A, as we discussed in Chapter 8. Nonetheless, partitioning the total variation in a response variable between that explained by the fixed factor and one or more nested random factors is a useful interpretative tool.

9.1.7 Assumptions for nested models

The assumptions of normality and homogeneity of within-cell variances apply to hypothesis tests in nested ANOVA models and they are checked using the same techniques (boxplots, mean vs variance plots and residuals vs mean plots) already described in Chapters 4 and 8. Traditionally, the observations within each cell (combination of main and nested factors) in the data set are used to check the assumptions. However, because the test of the main effect of A is based on the means of the levels of B when B is random, the normality and homogeneity of variance assumptions for the test of factor A apply to these means rather than within cell observations. You may, therefore, need to look at the assumptions separately for each hypothesis that uses a different denominator to make up the *F*-ratio. Transformations are applicable as usual (Chapter 4) but we know of no accepted non-parametric or robust (at least to unequal variances) tests specifically for nested designs. Any approach would require the main effect of A to be tested using the nested factor means as observations with one of the robust single factor tests described in Chapter 8. For non-normal data, the RT (rank transform, see Section 9.2.9) approach may also be useful, particularly when outliers are present. Of course, generalized linear models (GLMs; see Chapter 13) would also be applicable when the underlying distribution of the response variable is not normal but known to fit one from the exponential family suited to GLMs.

In many cases, it is the higher levels in the hierarchy that are of most interest. We would expect, from the Central Limit Theorem, that normality will be satisfied for all levels other than the lowest one in the hierarchy, because we are effectively working with means at higher levels. Means are more likely to be normally distributed, regardless of the underlying distribution of the observations.

The assumption of independence is also relevant for nested ANOVA models. The observations within each cell (e.g. level of B with A) are commonly measured at small spatial scales, such as quadrats within patches (Andrew & Underwood 1993). We need to design our study to ensure that these observations are independent of each other within each level of B.

9.1.8 Specific comparisons for nested designs

The logic and mechanics of planned and unplanned comparisons are the same as for single factor ANOVA models (Chapter 8) with two exceptions. First, we are usually only interested in comparisons between levels of factor A if it is fixed. The nested factors are commonly random so specific comparisons of levels of these factors within each level of the higher factor are rarely relevant. Second, we must use the appropriate standard error for comparisons of means of the fixed factor. The standard error for contrasts between A means should be based on $MS_{B(A)}$ if B is random, just as for the *F* test for factor A in the ANOVA model (see Table 9.4).

9.1.9 More complex designs

These designs can be extended to three or more nested factors (Table 9.6(a)) and are often used when there are multiple levels of subsampling, e.g. plants within treatments, pieces of tissue within each plant, sections cut from each piece of tissue, cells measured from each section. We have already described the study of Downes *et al.* (1993) who used three sites along a river, two riffles (shallow stony areas) at each site, with five groups of three stones within each riffle to examine hierarchical spatial variation in the distribution of stream invertebrates (Table 9.6(b)). Their linear model incorporated site, riffle nested within site, group nested within riffle within site and replicate stones within group within riffle within site:

$$(\text{density})_{ijkl} = \mu + (\text{site})_i + (\text{riffle within site})_{j(i)} + (\text{group within riffle within site})_{k(j(i))} + \varepsilon_{ijkl} \quad (9.8)$$

Another example is from Abrams *et al.* (1994), who examined variation in leaf structural parameters across three sites (xeric, mesic, wet-mesic) in Pennsylvania, with five or six different species at each site, six sapling trees of each species and replicate leaves from each tree. Their linear model incorporated site, species nested within site, tree nested within species within site and replicate measurements within tree within species within site:

$$(\text{leaf structure})_{ijkl} = \mu + (\text{site})_i + (\text{species within site})_{j(i)} + (\text{trees within species within site})_{k(j(i))} + \varepsilon_{ijkl} \quad (9.9)$$

Both Abrams *et al.* (1994) and Downes *et al.* (1993) calculated variance components for each factor (Table 9.6(b)). Since all nested factors were random in these studies, the *F*-ratio for the null hypothesis for each term in the model used the term immediately below as the denominator.

9.1.10 Design and power

If the main (highest) factor in a nested design is fixed, we could use formal power analysis based on specified and negotiated effect sizes (Chapters 7 and 8) to determine the number of groups nested within that main factor that we need to detect a particular treatment effect. If the nested factor B is random, then the power of the test for A will depend on the level of replication of B, and on the amount of variation among levels of B. For example, based on the experiment of Andrew & Underwood (1993) manipulating sea urchin densities, we would specify the desired effect size between density treatments and use an estimate of the variance between patches within each treatment to determine the number of patches required to achieve a given power. This simply becomes a single factor design using patch means so the methods outlined in Chapter 8 are appropriate.

This has implications for the design of nested experimental and sampling programs. The higher level "units" in nested designs are often increasingly costly, either because they are more expensive (e.g. whole animals vs pieces of tissue) or take longer to record (large spatial areas vs small quadrats). It is then tempting to take more replicates at lower levels in the design hierarchy. It is very important to realize that to increase the power of the test for fixed main effects, we need to increase the number of levels of the random factor immediately below the fixed factor. For example, Andrew & Underwood (1993) could improve the power of the test for differences in algal cover among sea urchin densities more by increasing the number of patches per treatment rather than the number of quadrats per patch. Nonetheless, smaller-scale noise as part of the apparent variation in factor B can still be important. From the expected mean squares for a two factor nested design (Table 9.4), we see that the $MS_{B(A)}$ includes two components, small-scale variation (σ_ε^2) and

Box 9.3 Calculations for optimal allocation of subsampling resources for two factor nested design based on Andrew & Underwood (1993)

Using the data from Andrew & Underwood (1993) as a pilot study and based on costs of 5 min to record a quadrat within a patch and 23 min to set up and work a patch (excluding quadrat recording time), we can estimate the optimal number of quadrats per patch:

$$n = \sqrt{\frac{C_{B(A)} s^2_{C(B(A))}}{C_{C(B(A))} s^2_{B(A)}}} = \sqrt{\frac{23 \times 298.60}{5 \times 1770.16}} = 0.88$$

Therefore, we would use a single quadrat per patch. If we set the total cost per density treatment at 4 h (240 min), we can determine the optimal number of patches per treatment if we have one quadrat per patch:

$$C_A = qC_{B(A)} + nqC_{C(B(A))}$$
$$240 = q \times 23 + 1 \times q \times 5$$
$$q = 8.57$$

The optimal experiment design would have nine patches per density treatment and one quadrat per patch.

the true variance between B groups ($n\sigma_\beta^2$). As we increase our subsampling effort (i.e. raise n), MS_B becomes increasingly dominated by σ_β^2. Therefore, while subsampling at levels below B has no direct effect on the power of the test of A, if there is considerable small-scale variation, then taking some replicates at lower levels will provide better variance estimates, and improve power.

At lower levels of nested designs, power is much less an issue, as degrees of freedom generally increase from top to bottom of hierarchical designs. Increases in replication at higher levels of the hierarchy will have cascading effects on power at lower levels. However, it must be remembered that formal power calculations would need to be done separately for each level, i.e., for each hypothesis of interest.

Note that power of tests of particular terms in a model may be increased by pooling non-significant terms with their error term, thus creating a pooled residual term with more degrees of freedom for tests of other terms. Issues and guidelines for pooling terms in multifactor ANOVA models will be discussed in Section 9.3.

Another important aspect of the design of studies that use a series of nested random factors is the allocation of limited resources to multiple spatial or temporal scales of sampling. For example, imagine we were following up the study of Andrew & Underwood (1993) who set up four sea urchin density treatment with four replicate patches within each treatment and five replicate quadrats within each patch. The number of treatments is obviously fixed, but in the new study, how should we allocate our sampling effort to the two different spatial levels in this design? Given limited resources, do we use more patches within each treatment, or more quadrats within each patch?

There are two criteria we use to decide on this relative allocation. First is the precision of the means for each level of the design or, conversely, the variance of these means. Second is the cost, in terms of money and/or time, of sampling each level in the design. We will illustrate the calculations for determining this relative allocation of resources for the study by Andrew & Underwood (1993) – see Box 9.3. This is a two factor nested design with p levels of A (density treatment), q levels of B (patches) nested within A (B(A)) and n replicate observations (quadrats) within each combination of density treatment and patch (C(B(A)), i.e. the Residual). Sokal & Rohlf (1995) illustrate the calculations for a three factor design. We use the variance components for each

random term in the model to estimate the variance associated with each term in the model separately from the other components of variation (Section 9.1.6). The costs (C) must also be determined, preferably from our pilot study where costs can be estimated empirically. The cost for each quadrat is simply the time and/or money required to place the quadrat and estimate the percentage cover of algae, say five minutes. The cost for each patch would be the time taken to move all the gear to each patch (20 minutes) and the time taken to move between quadrats in each patch (three minutes) but NOT the time taken to process a quadrat.

A number of textbooks (Snedecor & Cochran 1989, Sokal & Rohlf 1995, Underwood 1997) provide equations for relating costs and variances to determine the optimum number of replicates at each level of sampling (and see Andrew & Mapstone 1987). In a two factor design, the optimum number of replicates (e.g. quadrats) in each level of B (e.g. each patch) is:

$$n = \sqrt{\frac{C_{B(A)} s^2_{C(B(A))}}{C_{C(B(A))} s^2_{B(A)}}} \quad (9.10)$$

where C is the cost for the appropriate level and s^2 is the estimate of the variance, i.e. the mean square. Note that if the costs of recording a single quadrat are the same as the costs of setting up a new patch, then the sample size is just based on the ratio of the two variance components. Based on the variances and the costs listed above, the optimal number of quadrats per patch is 0.88, i.e. one (Box 9.3).

The number of patches (q) for each density treatment can be determined in two ways based on either the desired variance of the mean for each site (s_A^2) or the fixed total cost of sampling a site (C_A):

$$s_A^2 = \frac{ns^2_{B(A)} + s^2_{C(B(A))}}{nq} \quad (9.11)$$

$$C_A = qC_{B(A)} + nqC_{C(B(A))} \quad (9.12)$$

In the first case, we fix the desired level of precision for the mean of each site (s_A^2) and, using our values for n and the estimated variance components for quadrats and patches, solve for q. In the second case, we fix the total available cost for sampling each density and, again using our values for n and the estimated variance components for quadrats and patches, solve for q. In practice, having a fixed total cost, in time or money, is likely so the latter approach might be used more often. If we set the total cost for setting up each density treatment as four hours (240 minutes), then the number of patches would be 8.6, i.e. nine (Box 9.3). So based on these estimates, the most efficient design would be one quadrat per patch and nine quadrats per treatment. Note that these costs are guesses on our part so we are not suggesting that there was anything wrong with the design used by Andrew & Underwood (1993).

Keough & Mapstone (1995) made a number of sensible recommendations for deriving and using these values for sample size at each level of subsampling. First, the calculated sample sizes depend on the quality of the pilot data, particularly the variance estimates, and how well the variances in the subsequent main study will match those from the pilot study. It is important, therefore, that the pilot study is done in similar locations and at a similar time (e.g. season) to the main study. It is also important to check that these variance estimates still hold once the main research has started and adjust the sample sizes if necessary. It is much easier to reduce sample size during an ongoing research program than to increase them, so the initial sample sizes should be generous. Second, the sample size values will usually not be integers so they should be rounded up to the nearest integer. Finally, the calculations may recommend sample sizes of less than one, because the variance at that level is so small or the costs so cheap. However, some level of replication is necessary for sensible inference and, remembering that pilot studies may underestimate the true variance, we recommend that more than one replicate at any level should always be used.

9.2 Factorial designs

An alternative multifactor linear model is used when our design incorporates two or more factors that are crossed with each other. The term crossed indicates that all combinations of the factors are

Table 9.7 Illustration of marginal and cell means for a two factor factorial ANOVA design. Data from Quinn (1988) where factor A is limpet density, factor B is season and the response variable is number of egg masses per limpet in three replicate enclosures per cell

	B_1	B_2	B_j	Marginal means A
A_1	μ_{11}	μ_{12}	μ_{1j}	$\mu_{i=1}$
A_2	μ_{21}	μ_{22}	μ_{2j}	$\mu_{i=2}$
A_i	μ_{i1}	μ_{i2}	μ_{ij}	μ_i
Marginal means B	$\mu_{j=1}$	$\mu_{j=2}$	μ_j	Grand mean μ

	Factor B (B_j) Season	B_1 Spring	B_2 Summer	Factor A marginal means
Factor A (A_i)	Density			
A_1	8	$\bar{y}_{11} = 2.417$	$\bar{y}_{12} = 1.833$	$\bar{y}_{i=1} = 2.125$
A_2	15	$\bar{y}_{21} = 2.177$	$\bar{y}_{22} = 1.178$	$\bar{y}_{i=2} = 1.677$
A_3	30	$\bar{y}_{31} = 1.565$	$\bar{y}_{32} = 0.811$	$\bar{y}_{i=3} = 1.188$
A_4	45	$\bar{y}_{41} = 1.200$	$\bar{y}_{42} = 0.593$	$\bar{y}_{i=4} = 0.896$
Factor B marginal means		$\bar{y}_{j=1} = 1.840$	$\bar{y}_{j=2} = 1.104$	$\bar{y} = 1.472$

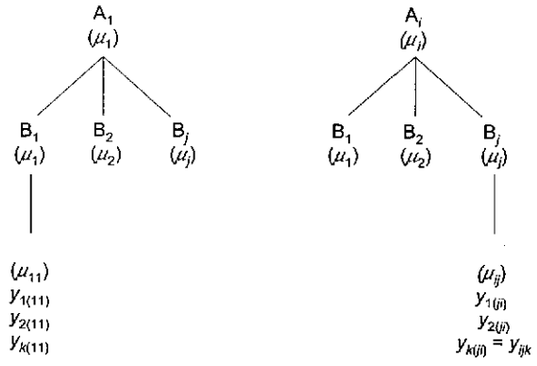

Figure 9.2 Part of data set for two factor crossed ANOVA, with p levels of factor A ($i = 1$ to p), q levels of factor B ($j = 1$ to q), where the levels of B are the same and crossed with each level of A, and n replicate observations within each combination (cell) of A and B ($k = 1$ to n).

included in the design and that every level (group) of each factor occurs in combination with every level of the other factors. Such designs are also termed factorial. This pattern is in contrast to nested designs, where the levels of the nested factor are different within each level of the main factor. We will first consider factorial (crossed) designs with two factors, where every level of one factor occurs at every level of the other factor and both factors are of equal importance – see Figure 9.2 and Table 9.7.

Factorial designs are most often used for manipulative experiments. For example, Poulson & Platt (1996) examined the effects of light microenvironment (three levels: beneath canopy, single treefall gap, multiple treefall gap) and seedling height class (three levels: 1–2 m small, 2–4 m medium, 4–8 m large) on the difference in growth between sugar maple and beech saplings (the response variable was the difference in growth of paired seedlings of each species). There were five replicate seedling pairs for each of the nine microenvironment–height combinations. Another example comes from Maret & Collins (1996), who set up an experiment to test the effects of invertebrate food level and the presence or absence of tadpoles on variation in size among larval salamanders. There were two factors: two levels of ration of invertebrate prey (low and high amounts of brine shrimp per day) and two levels of tadpole supplementation (with and without). There were originally eight replicate aquaria in each of the four cells, although some aquaria were omitted from analysis because one or more salamander larvae died. The response variable was mean snout–vent length of salamanders in each aquarium.

In these two examples, both factors in the design are fixed, i.e. all possible levels of interest for the two factors have been used in the study and our inference is restricted to these levels. These are analyzed with fixed effects linear models, also termed Model 1 analyses of variance.

Factorial designs can include random factors that are often randomly chosen spatial or temporal units. Designs that include only random factors are analyzed with random effects models, termed Model 2 analyses of variance, although these are unusual in biology. One example is from Kause *et al.* (1999), who examined phenotypic plasticity in the foraging behavior of sawfly larvae with an experiment that used six species of sawflies and 20 individual mountain birch trees that represented a range of leaf qualities for the herbivorous sawfly larvae. There were between four and six larvae per tree and species combination and the response variable was an aspect of foraging behavior (e.g. number of meals, relative consumption rate etc.). Both sawfly species and individual tree were random factors as they were a sample from all possible herbivorous sawflies and all possible trees.

Designs with a combination of fixed and random factors are analyzed with mixed linear models, also termed Model 3 analyses of variance. Including a random factor in a multifactor design is important in biology, because it allows us to generalize the effects of a fixed factor to the population of spatial or temporal units (Beck 1997). For example, Brunet (1996) tested the effects of position on an inflorescence and randomly chosen plants on fruit and seed production of a perennial herb. This was a two factor design with flower position as the fixed factor and individual plant as the random factor. A second example comes from Twombly (1996), who randomly assigned copepod nauplii from 15 sibships to one of four food treatments (high constant food and high switched to low at three different naupliar stages); there were four replicate dishes (each containing two nauplii) per factor combination and the response variable was age at metamorphosis. Food treatment was a fixed factor and sibship was a random factor.

Factorial designs can include three or more factors (Section 9.2.12), although we will illustrate the principles based on two factor designs. Factorial designs allow us to measure two different sorts of factor effects.

1. The main effect of each factor is the effect of each factor independent of (pooling over) the other factors.

2. The interaction between factors is a measure of how the effects of one factor depend on the level of one or more additional factors. The absence of an interaction means that the combined effect of two or more factors is predictable by just adding their individual effects together. The presence of an interaction indicates a synergistic or antagonistic effect of the two factors.

We can only measure interaction effects in factorial (crossed) designs. In nested designs where factor B is nested within factor A, different levels of B are used in each level of A so any interaction between A and B cannot be assessed. When all possible combinations of the two (or more) factors are used in factorial designs they are called complete factorials. Sometimes this is logistically impossible because the experiment would be too big and/or costly, so a subset of factor combinations is used and the design is termed a fractional factorial. Such designs are more difficult to analyze because not all interactions can be measured – see Section 9.2.12.

Fecundity of limpets: effects of season and adult density

Our first worked example of a factorial ANOVA design and analysis is from Quinn (1988). He examined the effects of season (two levels, winter/spring and summer/autumn) and adult density (four levels, 8, 15, 30 and 45 animals per 225 cm^2) on the production of egg masses by rocky intertidal pulmonate limpets (*Siphonaria diemenensis*). Limpets (approx. 10 mm shell length) were enclosed in 225 cm^2 stainless steel mesh enclosures attached to the rocky platform. There were eight treatment combinations (four densities at each of two seasons) and three replicate enclosures per treatment combination. Note that all four densities were used in both seasons, hence a factorial or crossed design. One of the important questions being asked with this experiment was whether the effect of density on number of egg masses per limpet depended on season. Quinn (1988) predicted that the density effect would be greater in summer/autumn, when algal food was

scarce, than in winter/spring, when algal food was more abundant.

Quinn (1988) described another experiment looking at the same species of limpet lower on the shore. Here the limpets were bigger (15–20 mm shell length) and there was much less seasonal variation in the availability of algal food, algal cover being high all year round. The same two factors were used for this experiment but only three densities were included: 6, 12 and 24 limpets per 225 cm^2. So there were six treatment combinations (three densities at each of two seasons) and three replicate enclosures per treatment combination. The analyses of both experiments are in Box 9.4.

Oysters, limpets and mangrove forests

Our second example is from Minchinton & Ross (1999), who examined the distribution of oysters, and their suitability as habitat for limpets in a temperate mangrove forest. They chose two sites about 600 m apart and at each site recorded the density of oysters in four zones running up the shore: seaward zone without mangrove trees, seaward zone with mangrove trees, middle zone with trees, and a landward zone at the upper levels. In each of the eight combinations of site and zone, they used five quadrats to sample oysters (response variable) on the forest floor. An additional study examined the distribution of limpets on oysters on bent mangrove tree trunks. They used two sites, three zones (obviously the seaward zone without trees was not included) and two orientations of mangrove trunk (upper facing canopy and lower facing forest floor). This was a three factor sampling design with five quadrats in each of the 12 cells and densities of limpets per oyster surface as the response variable. For both designs, site was a random factor, representing all possible sites within the mangrove forest, and zone and orientation were fixed factors. The analyses of these data are in Box 9.5.

Box 9.4 Worked example of two factor fixed effects ANOVA

Quinn (1988) examined the effects of season (winter/spring and summer/autumn) and adult density (8, 15, 30 and 45 animals per 225 cm^2 enclosure) on the production of egg masses by intertidal pulmonate limpets (*Siphonaria diemenensis*). There were three replicate enclosures per treatment combination and the response variable was the number of egg masses per limpet in each enclosure.

The null hypotheses were as follows.

No difference between mean number of egg masses laid in each season, pooling densities.

No difference in mean number of egg masses laid at each density, pooling seasons.

No interaction between season and density, i.e. the effect of density on mean numbers of egg masses laid is independent of season and vice versa.

Source	df	MS	F	P
Density	3	1.76	9.67	0.001
Linear	1	5.02	27.58	<0.001
Quadratic	1	0.24	1.29	0.272
Season	1	3.25	17.84	0.001
Density × season	3	0.06	0.30	0.824
Residual	16	0.18		

There were no outliers and the residual plot (Figure 9.4(a)) did not suggest problems with assumptions. There was no evidence of an interaction ($P = 0.824$, see Figure 9.5(a)). There were significant effects of season (more egg masses in winter/spring than summer/autumn) and density. The main effect of density was further analyzed with orthogonal polynomials (see Chapter 8 and Section 9.2.10). There was a significant negative linear trend in egg mass production with density but no quadratic trend.

Quinn (1988) did a similar experiment at a lower level of the same shore where the limpets were larger. Different densities were used (6, 12, 24) but the same two seasons with three replicate enclosures per treatment combination. The null hypotheses were the same as above, except that there were only three densities. Again, the residual plot did not suggest any problem with variance heterogeneity (Figure 9.4(b)).

Source	df	MS	F	P
Density	2	2.00	13.98	0.001
Season	1	17.15	119.85	<0.001
Density × season	2	0.85	5.91	0.016
Density 6 vs 12 & 24 × season	1	1.53	10.66	0.007
Linear density × season	1	1.44	10.07	0.008
Residual	12	0.14		

There was a significant interaction between density and season ($P = 0.016$, Figure 9.5(b)). Treatment–contrast interaction tests showed that the comparison between control density and increased density varied between seasons and the linear trend in density was also significantly different between seasons. We also tested simple main effects of density separately for each season.

Source	df	MS	F	P
Winter density	2	0.17	1.21	0.331
Summer density	2	2.67	18.69	<0.001
Residual	12	0.14		

The effect of density was only significant in summer, not in winter. Note that the original $MS_{Residual}$ was used for both tests.

Designs with more than two factors will be examined in Section 9.2.12. A two factor factorial design is illustrated in Figure 9.2 with a factor relationship diagram. Factor A has p groups ($i = 1$ to p), factor B has q groups ($j = 1$ to q) crossed with each level of A and there are n_i replicates ($k = 1$ to n_i) within each combination of A and B categories, i.e. each cell. Note that every level of factor B is crossed with every level of factor A and vice versa. For the moment, assume the number of replicate observations (n) in each combination of A and B is the same. Unequal sample sizes will be discussed in Section 9.2.6. There will be a total of pq cells in this factorial design with n replicate observations in each cell. From Quinn (1988), p was four limpet density treatments (factor A), q was two seasons (factor B) and n was three enclosures within each cell. From Minchinton & Ross (1999), p was four zones (factor A), q was two sites (factor B) and n was five quadrats within each cell.

We need to distinguish between two types of means in multifactor crossed designs (Table 9.7).

- Marginal means are the means for the levels of one factor pooling over the levels of the

9.2.1 Linear models for factorial designs

In the sections that follow, we will describe two factor designs and their associated linear models.

Box 9.5 | Worked example of two factor mixed effects ANOVA

Minchinton & Ross (1999) examined the distribution of oysters, and their suitability as habitat for limpets in a temperate mangrove forest. There were two factors: randomly chosen sites (two sites about 600 m apart) and fixed zones (four levels running up the shore: seaward zone without mangrove trees, seaward zone with mangrove trees, middle zone with trees, and a landward zone at the upper levels). In each of the eight combinations of site and zone, they used five quadrats to sample limpets on oyster shells (response variable) on the forest floor. There was a strong relationship between cell means and cell variances (Figure 9.6), indicating that number of limpets was positively skewed. After transformation to square roots (×100, representing limpets per 100 oyster shells), much of the mean–variance relationship was removed, indicating that the distribution of the response variable was more symmetrical. Like Minchinton & Ross (1999), we analyzed the transformed variable.

The null hypotheses were as follows.

No difference in the mean square root number of limpets per quadrat between zones, pooling across all possible sites.

No difference in the mean square root number of limpets per quadrat between all possible sites, pooling across zones.

No interaction between zone and site, i.e. the effect of zone on the mean square root number of limpets per quadrat is independent of all possible sites that could have been used and vice versa.

The two factor mixed model ANOVA tested the fixed effect of zone against the interaction term, with only 3 and 3 df, because site was random.

Source	df	MS	*F*	*P*	Variance component	%
Zone	3	13.08	1.24	0.433	(0.25)	
Site	1	6.37	1.84	0.184	0.15	2.90
Zone × site	3	10.59	3.06	0.042	1.43	28.36
Residual	32	3.46			3.46	68.74

The H_0 of no interaction between zone and site was rejected, indicating that the effect of zone was not consistent between sites in this mangrove forest. This is clear in Figure 9.7 where site A has fewest limpets in the middle zone whereas site B has the most limpets in this zone. Most of the variance in limpet densities was unexplained, although the interaction explained nearly ten times more than the main effect of site.

Note that the *F*-ratio for zone would have been 3.78 with 3 and 32 df ($P = 0.020$) if site had been considered fixed, resulting in rejection of the H_0 of no effect of zone. We would be more confident of a zone effect for just the two sites used (site fixed), than a zone effect for all possible sites we could have used (site random).

second factor, so the marginal mean A_1 is the mean for the first level of A pooling over the levels of B. For example, the marginal mean for density eight from Quinn (1988) is the mean number of egg masses per limpet from all possible enclosures with eight limpets, pooling both seasons. The marginal mean for each level of A is μ_i and the marginal mean for each level of B is μ_j.

- Cell means are the means of the observations within each combination of A and B. For example, the mean number of egg masses per limpet from enclosures within each density–season combination. The cell means for each combination of A and B are μ_{ij}.

Model 1 – both factors fixed

The linear ANOVA model for a factorial design with two fixed factors is an extension of the model used for single factor designs in Chapter 8. The two factor effects model is:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \quad (9.13)$$

Statistical details of the crossed ANOVA model, including estimation of its parameters, are provided in Box 9.6.

Box 9.6 | The fixed effects factorial ANOVA model and its parameters

The linear ANOVA models for a factorial design with two fixed factors are extensions of the models used for single factor designs in Chapter 8. The effects model is:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

In model 9.13:

y_{ijk} is the *k*th replicate observation from the combination of the *i*th level of factor A and *j*th level of factor B, i.e. cell *ij*.

μ is the overall (constant) population mean of the response variable.

α_i is effect of *i*th level factor A, pooling the levels of factor B. This is the main effect of factor A, the effect of A pooling (independent of) factor B, and is defined as the difference between each A marginal mean and the overall mean $(\mu_i - \mu)$.

β_j is effect of *j*th level of factor B, pooling the levels of factor A, which is the difference between each B marginal mean and the overall mean. This is the main effect of factor B, the effect of B pooling (independent of) factor A, and is defined as the difference between each B marginal mean and the overall mean $(\mu_j - \mu)$.

$(\alpha\beta)_{ij}$ is the effect of the interaction of the *i*th level of A and the *j*th level of B and is defined as $(\mu_{ij} - \mu_i - \mu_j + \mu)$. Interactions measure whether the effect of one factor depends on the levels of the other factor and vice versa. This can also be viewed as measuring whether the effects of one factor are independent of the other second factor.

ε_{ijk} is random or unexplained error associated with the *k*th replicate observation from the combination of the *i*th level of factor A and *j*th level of factor B. These error terms are assumed to be normally distributed at each combination of factor levels, with a mean of zero $[E(\varepsilon_{ij}) = 0]$ and a variance of σ_ε^2.

This fixed effects model is overparameterized because the number of means (combinations of factors plus overall mean) is less than the number of model

parameters to be estimated ($\mu, \alpha_1, \alpha_2, \ldots, \beta_1, \beta_2, \ldots, (\alpha\beta)_{11}, (\alpha\beta)_{12}, \ldots$). Overcoming this problem so we can estimate model parameters requires a series of "sum-to-zero" constraints:

$$\sum_{i=1}^{p} \alpha_i = 0, \sum_{j=1}^{q} \beta_j = 0, \sum_{i=1}^{p} (\alpha\beta)_{ij} = 0, \sum_{j=1}^{q} (\alpha\beta)_{ij} = 0.$$

These constraints appear formidable but simply imply that the sum of the effects of factor A, pooling B, and the sum of the effects of factor B, pooling A, are both zero. Additionally, the sum of the interaction effects for each level A and for each level of B are also zero. These constraints are necessary for fitting effects models, although such constraints have been criticized (Chapter 8), and further technical discussion of this issue can be found in Hocking (1996), Searle (1993) and Yandell (1997).

An alternative to imposing constraints on the effects model is to fit a much simpler means model:

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

where μ_{ij} is the mean of cell ij and ε_{ijk} is random or unexplained variation. The means model basically treats the analysis as a large single factor ANOVA comparing all cells and tests specific hypotheses about interactions and main effects. The means model estimates A and B means by averaging the cell means across rows or columns (Searle 1993), so it has certain advantages for unbalanced designs by ignoring the sample sizes completely. Means models are mainly useful for missing cells designs (see Section 9.2.6).

Estimating the parameters of the factorial linear model 9.13 follows the methods outlined for a single factor model in Chapter 8 and nested models in Box 9.2 with the added complication of estimating interaction effects. Cell means (μ_{ij}) for each combination of A and B are estimated from the sample mean of the observations in each cell, based on the sample size of the particular cell if sample sizes are unequal.

The factor level (marginal) mean for each level of A pooling levels of B is simply the mean of the sample means for each cell at level i of factor A, averaged across the levels of B (Table 9.2). An analogous calculation can be done for factor B means. These are unweighted means and ignore any difference in sample sizes between cells.

An alternative approach is to calculate a weighted marginal mean, which averages the observations for each level of A taking into account different n_{ij} within each cell. If we have a fully balanced design (all n_{ij} equal), then the unweighted and weighted estimates of factor level means will obviously be the same. If we have unequal numbers of observations per cell (some n_{ij} different), then the estimates will be different. In unbalanced crossed designs, only Type III SS are based on unweighted marginal means and therefore only F-ratio statistics based on Type III SS test hypotheses about unweighted marginal means (Section 9.2.6). Our preference for unbalanced designs is to estimate and test hypotheses about unweighted means.

Standard errors for these means are based on the mean squares used in the denominator of the appropriate F test of the H_0 that the population means are equal (Table 9.11). Note that the OLS standard error for a specific mean will be different from that calculated if we treat the observations producing that mean as a single sample and calculate the standard error as described in Chapter 2. The former uses a pooled variance estimate for the whole data set whereas the latter only uses the variance of the observations producing the mean.

The estimates of α_i ($\mu_i - \mu$) and β_j ($\mu_j - \mu$) are the differences between the mean of each A level or each B level and the overall mean, $\bar{y}_i - \bar{y}$ and $\bar{y}_j - \bar{y}$ respectively. Interaction effects measure how much the effect of one factor depends on the level of the other factor and vice versa. If there was no interaction between the two factors, we would expect the cell means to be represented by the sum of the overall mean and the main effects:

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

Therefore, the effect of the interaction between the ith level of A and jth level of B $(\alpha\beta)_{ij}$ can be defined as the difference between the ijth cell mean and its value we would expect if there was no interaction:

$$\alpha\beta_{ij} = \mu_{ij} - \mu_i - \mu_j + \mu$$

which is estimated by:

$$\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y}$$

This represents those effects not due to the overall mean and the main effects.

Note that in practice biologists rarely calculate the estimated factor or interaction effects, instead focusing on contrasts of marginal or cell means. The exception is when we have random factors in our model and estimating variance components is often of interest.

Using the example from Quinn (1988):

$$\begin{aligned}(\text{no. egg masses per limpet})_{ijk} &= \mu + \\ (\text{effect of density})_i &+ (\text{effect of season})_j + \\ (\text{interaction between density and season})_{ij} &+ \\ \varepsilon_{ijk} & \qquad (9.14)\end{aligned}$$

In models 9.13 and 9.14 we have the following:

y_{ijk} is the number of egg masses per limpet from the kth replicate enclosure from the combination of the ith density and jth season, i.e. cell ij.

μ is the overall (constant) population mean number of egg masses per limpet from all possible enclosures in the eight density–season combinations.

α_i is the main effect of ith density on the number of egg masses per limpet, pooling (independent of) seasons.

β_j is the main effect of jth season on the number of egg masses per limpet, pooling (independent of) densities.

$(\alpha\beta)_{ij}$ is the effect on the number of egg masses per limpet of the interaction of the ith density and jth season. This interaction measures whether the effect of density on number of egg masses per limpet depends on season and vice versa, also whether the effect of density is independent of the effect of season.

ε_{ijk} is random or unexplained error associated with the kth replicate enclosure from the combination of the ith level of density and jth level of season. This measures the random error associated with the number of egg masses per limpet in each enclosure and the existence of this error is why replicates within each cell produce different values for the response variable.

Model 2 – both factors random

These designs are relatively uncommon in biological research (but see Kause *et al.* (1999) for a recent example) so we will not examine them in detail – see Neter *et al.* (1996).

Model 3 – one factor fixed and one random

The linear model for a factorial design with one fixed and one random factor is the same as

outlined in 9.13 although the interpretation of the terms is different. Using the example from Minchinton & Ross (1999):

$$\begin{aligned}&(\text{density of oysters})_{ijk} = \mu + \\ &(\text{effect of intertidal zone})_i + \\ &(\text{effect of randomly chosen site})_j + \\ &(\text{interaction between zone and site})_{ij} + \varepsilon_{ijk}\end{aligned} \quad (9.15)$$

In models 9.13 and 9.15 we find the following.

y_{ijk} is the density of oysters from the kth quadrat from the combination of the ith zone and jth site.

μ is the overall (constant) population mean density of oysters.

α_i is effect of the ith zone on the density of oysters, pooling all possible sites.

β_j is a random variable with a mean of zero and a variance of σ^2_β, measuring the variance in mean density of oysters across all possible sites that could have been used, pooling zones.

$(\alpha\beta)_{ij}$ is a random variable with a mean of zero and a variance of $\sigma^2_{\alpha\beta}$ measuring the variance of the interaction between zone and site across all possible sites that could have been used. Biologically, this interaction term measures whether the zone effect is consistent across all possible randomly chosen sites.

ε_{ijk} is random or unexplained error associated with the kth replicate quadrat from the combination of the ith level of zone and jth level of site. This measures the random error associated with the density of oysters in each quadrat.

Predicted values and residuals

If we replace the parameters in our model by their OLS estimates (Box 9.6), it turns out that the predicted or fitted values of the response variable from our linear model (9.13) are:

$$\hat{y}_{ijk} = \bar{y} + (\bar{y}_i - \bar{y}) + (\bar{y}_j - \bar{y}) + (\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y}) = \bar{y}_{ij} \quad (9.16)$$

So any predicted Y-value is predicted by the sample mean for the cell that contains the Y-value. For example, the predicted number of egg masses per limpet for enclosure one in spring for the density of eight limpets is the sample cell mean for spring for the density of eight limpets.

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

$$e_{ijk} = y_{ijk} - \bar{y}_{ij} \quad (9.17)$$

For example, the residuals from the model relating number of egg masses per limpet to limpet density, season and their interaction are the differences between the observed number of egg masses per limpet in each enclosure and the mean number of egg masses per limpet from the enclosures within each limpet density and season combination that contained that enclosure. Note that the sum of the residuals within each cell ($\Sigma_{k=1}^{n} e_{ijk}$) equals zero. As in all linear models, 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 9.2.8). The OLS estimate of σ_ε^2 is the sample variance of these residuals and is termed the Residual (or Error) mean square and is calculated as part of the partitioning of the total variation in the response variable described in the next section.

9.2.2 Analysis of variance

The ANOVA table for a two factor factorial design with equal sample sizes per cell is shown in Table 9.8. The SS_A measures the sum of squared differences between each A marginal mean and the overall mean; the SS_B measures the sum of squared differences between each B marginal mean and the overall mean; the SS_{AB} measures the sum of squared differences for a particular contrast involving cell means, marginal means and the overall mean; $SS_{Residual}$ measures the difference between each replicate observation and the appropriate cell mean, summed across all cells. These SS represent an additive partitioning of the total SS in the response variable:

$$SS_{Total} = SS_A + SS_B + SS_{AB} + SS_{Residual} \quad (9.18)$$

In unbalanced designs (unequal n), there is no simple additive partitioning of the SS_{Total}, which causes some difficulties in the ANOVA. There are three different ways of determining the SS that represent very different philosophies for handling unequal sample sizes and we will discuss these in Section 9.2.6.

The degrees of freedom are calculated as usual (the number of components making up the

Table 9.8 | ANOVA table for two factor crossed model

Source	SS	df	MS
A	$nq\sum_{i=1}^{p}(\bar{y}_i - \bar{y})^2$	$p-1$	$\frac{SS_A}{p-1}$
B	$np\sum_{j=1}^{q}(\bar{y}_j - \bar{y})^2$	$q-1$	$\frac{SS_B}{q-1}$
AB	$n\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y})^2$	$(p-1)(q-1)$	$\frac{SS_{AB}}{(p-1)(q-1)}$
Residual	$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij})^2$	$pq(n-1)$	$\frac{SS_{Residual}}{(pq(n-1)}$
Total	$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(y_{ijk} - \bar{y})^2$	$pqn-1$	

Table 9.9 | Expected mean squares for a two factor crossed ANOVA model with both factors fixed (Model 1) or random (Model 2)

	A, B fixed	A, B random
MS_A	$\sigma_\varepsilon^2 + nq\frac{\sum_{i=1}^{p}\alpha_i^2}{p-1}$	$\sigma_\varepsilon^2 + n\sigma^2_{\alpha\beta} + nq\sigma_\alpha^2$
MS_B	$\sigma_\varepsilon^2 + np\frac{\sum_{j=1}^{q}\beta_j^2}{q-1}$	$\sigma_\varepsilon^2 + n\sigma^2_{\alpha\beta} + np\sigma_\beta^2$
MS_{AB}	$\sigma_\varepsilon^2 + n\frac{\sum_{i=1}^{p}\sum_{j=1}^{q}(\alpha\beta)_{ij}^2}{(p-1)(q-1)}$	$\sigma_\varepsilon^2 + n\sigma^2_{\alpha\beta}$
$MS_{Residual}$	σ_ε^2	σ_ε^2

variance minus one), with the df_{AB} being a product of the df_A and df_B.

The SS are divided by the df to produce mean squares (MS or variances) as we have done previously for single factor and nested ANOVA models. Statisticians have determined what population values these sample mean squares estimate, i.e. what their expected values are, if the assumption of homogeneity of variance holds (Table 9.9, Table 9.10). For all three models (fixed effects, random effects and mixed effects), the $MS_{Residual}$ estimates σ_ε^2 (the variation in the error terms in each cell, pooled across all cells). The expected values for MS_A, MS_B and MS_{AB} depend critically on whether the factors are fixed or random. When both factors A and B are fixed, the mean squares estimate the residual variance plus a measure of the fixed factor or interaction effects. When both factors are random, MS_{AB} estimates the residual variance plus the added variance due to the interaction

Table 9.10 Expected mean squares for a two versions of a two factor crossed mixed ANOVA model (Model 3: A fixed, B random): restricted version imposes constraints on interaction terms and unrestricted imposes no such constraints

	Restricted version	Unrestricted version
MS_A	$\sigma_\varepsilon^2 + n\sigma_{\alpha\beta}^2 + nq\frac{\sum_{i=1}^{p}\alpha_i^2}{p-1}$	$\sigma_\varepsilon^2 + n\sigma_{\alpha\beta}^2 + nq\frac{\sum_{i=1}^{p}\alpha_i^2}{p-1}$
MS_B	$\sigma_\varepsilon^2 + np\sigma_\beta^2$	$\sigma_\varepsilon^2 + n\sigma_{\alpha\beta}^2 + np\sigma_\beta^2$
MS_{AB}	$\sigma_\varepsilon^2 + n\sigma_{\alpha\beta}^2$	$\sigma_\varepsilon^2 + n\sigma_{\alpha\beta}^2$
$MS_{Residual}$	σ_ε^2	σ_ε^2

terms, the mean squares for the main effects estimate the residual variance plus the added variance due to the interaction terms plus the added variance due to the random main effect of the relevant factor.

Things get even messier when we have a mixed model (A fixed, B random). The two alternative approaches for mixed models for factorial ANOVAs described in Box 9.7 produce different expected mean squares (Table 9.10), and the choice between the two versions of the mixed model has created much discussion in the statistical literature (Box 9.7). We recommend Model I, which results in MS_{AB} estimating the residual variance plus the added variance due to the interaction terms, MS_B estimating the residual variance plus the added variance due to the random main effect of B and MS_A estimating the residual variance plus the added variance due to the interaction terms plus the fixed factor A effects. The expectation for the mean square of the fixed factor in a two factor mixed model includes three components: the residual variance, the interaction variance and fixed factor effects.

As we will see in the next section, the different approaches to determining expected mean squares in ANOVA models have critical implications for the construction of hypothesis tests. The expected values for mean squares in three factor random and mixed models are even more complicated but following the Model I approach, the same general principles apply. Expected mean squares for fixed factors and their interactions will include terms for variance due to higher order random interactions (Table 9.10). Many texts provide algorithms for calculating expected mean squares for any number and combination of fixed and random factors (e.g. Neter *et al.* 1996, Underwood 1997, Winer *et al.* 1991).

9.2.3 Null hypotheses

There are three general H_0s we can test in a two factor factorial ANOVA. The first two are tests of main effects and the third is the test of the interaction. The specific null hypotheses being tested in a factorial linear model depend on whether the factors are fixed or random (see Box 9.8 for terminology).

Fixed effects models

FACTOR A

$H_0(A)$: $\mu_1 = \mu_2 = \ldots = \mu_i = \ldots = \mu_p$. This H_0 states there is no difference between the marginal means for factor A pooling over the levels of factor B (Table 9.7). For example, no difference in the mean number of egg masses per limpet for each level of density, pooling over the two seasons (Quinn 1988). This is equivalent to $H_0(A)$: $\alpha_1 = \alpha_2 = \ldots = \alpha_i = 0$, i.e. no effect of any level of factor A pooling over the levels of factor B. For example, no effect of any of the four densities on the mean number of egg masses per limpet, pooling the two seasons.

Box 9.7 The mixed factorial model and the mixed models controversy

Return to model 9.13 for a factorial design with two factors:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

When one of the factors, such as B, is random then two modifications occur. First, β_j is a random variable with a mean of zero and, most importantly, a variance of σ_β^2 measuring the variance in mean values of the response variable across all the possible levels of factor B that could have been used. Second, $(\alpha\beta)_{ij}$ is a random variable with a mean of zero and a variance of $\sigma_{\alpha\beta}^2$ (well, strictly $[(p-1)/p]\sigma_{\alpha\beta}^2$ to simplify the expected values of the mean squares – see Neter *et al.* 1996) measuring the variance across all the possible interaction terms. This interaction term measures whether the fixed effect of A is consistent across all possible randomly chosen levels of B.

To estimate the parameters of this model, we impose two sum-to-zero constraints. The first implies that the sum of the effects of the fixed factor A, pooling B, is zero and is the same as we used for the fixed effects model.

$$\sum_{i=1}^{p}\alpha_i = 0$$

The second implies that the sum of the interaction effects across the levels of A is also zero:

$$\sum_{i=1}^{p}(\alpha\beta)_{ij} = 0$$

This constraint also defines a covariance between pairs of interaction terms within each level of factor B, i.e. any two interaction terms will not be independent within each level of B. Using the Minchinton & Ross (1999) example, this model allows for the limpet densities per quadrat within a site to be positively or negatively correlated. The version of the mixed model that imposes this constraint originates with Scheffé (1959) and is termed the restricted (or Σ-restricted) model (Neter *et al.* 1996, Searle *et al.* 1992), also Model I in Ayres & Thomas (1990) and the constrained parameters (CP) model (Voss 1999), and is the version most commonly presented in linear models texts.

An alternative model (Model II) is one that does not impose any restrictions on the interaction terms and is termed, not surprisingly, the unrestricted model or the unconstrained parameters model (Voss 1999). This model implies that any two interaction terms are independent, within each level of A and B, and is recommended by a number of influential authors, including Hocking (1996), Milliken & Johnson (1984) and Searle *et al.* (1992). Using the Minchinton & Ross (1999) example, this model assumes that the covariance of limpet densities per quadrat within a site is the same for each pair of zones.

The two approaches for mixed models result in different expected mean squares (Table 9.10). Model I results in MS_{AB} estimating the residual variance plus the added variance due to the interaction terms, MS_B estimating the residual variance plus the added variance due to the random main effect of B and MS_A estimating the residual variance plus the added variance due to the interaction terms plus

the fixed factor A effects. The alternative approach (Model II) results in a different expectation for MS_B, which now estimates the residual variance plus the variance due to the interaction and the variance due to the random main effect of B. Note that the difference in the two approaches is only in the expectation for the mean square for the random factor, not the fixed factor or the interaction.

The expected mean squares in Table 9.10 indicate that the test of the random factor B will be different under the two models because the expectation of MS_B changes. Under the unrestricted Model II, B should be tested against the MS_{AB}, in contrast to Model I where it is tested against the $MS_{Residual}$. Which version of the mixed model is most appropriate for testing main effects of factor B has been an issue of considerable debate among statisticians (Hocking 1985, Schwarz 1993, Searle *et al.* 1992, Voss 1999) and among biologists (Ayres & Thomas 1990, Fry 1992), although the discussion will be difficult for most biologists to appreciate as it involves a reasonably high level of statistical detail. Ayres & Thomas (1990) argued that the covariance assumptions behind Model II (i.e. independent interaction effects) need to be carefully assessed before it could be applied (but see also Fry 1992). It is difficult to determine, in most cases, whether biological data are likely to meet the assumption of completely independent interaction terms.

Voss (1999) proposed that the test for factor B based on Model I is correct no matter which of the two alternative formulations for expected mean squares are used for the mixed model. He argued that the H_0 for no main effects of factor B in Model II is actually that $\sigma_{\alpha\beta}^2 = \sigma_\beta^2 = 0$ which results in the same *F*-ratio test as in Model I. Voss (1999) claimed that this effectively resolved the controversy over expected mean squares for random factors in mixed models and their subsequent hypothesis tests.

In a more radical approach, Nelder & Lane (1995) proposed that usual sum-to-zero constraints imposed when using overparameterized effects models (Box 9.6) are unnecessary and pointed out that if we don't apply such constraints, the expected mean squares for factors A and B both include the effect of the interaction. Indeed, the expected mean squares, and *F*-ratios for hypothesis tests, for each term become basically identical for all combinations of fixed and random factors. Under this model for expected mean squares, which is not conventional, testing fixed main effects is relevant even in the presence of interactions because we are testing for the effect of the fixed factor over and above the interaction. Expected mean squares and appropriate hypothesis tests in factorial ANOVA models are obviously still a topic of research and debate among statisticians.

FACTOR B

H_0(B): $\mu_1 = \mu_2 = \ldots = \mu_j = \ldots = \mu_q$. This H_0 states there is no difference between the marginal means for factor B pooling over the levels of factor A (Table 9.7). For example, no difference in the mean number of egg masses per limpet for each level of season, pooling over the four densities (Quinn 1988). This is equivalent to H_0(B): $\beta_1 = \beta_2 = \ldots = \beta_j = 0$, i.e. no effect of any level of factor B pooling over the levels of factor A. For example, no effect of either of the two seasons on the mean number of egg masses per limpet, pooling the four densities.

INTERACTION BETWEEN A AND B

H_0(AB): $\mu_{ij} - \mu_i - \mu_j + \mu = 0$ for all levels of A and all levels of B. This is testing that there are no effects in addition to the overall mean and the main effects. For example, there are no effects on the mean number of egg masses per limpet besides the main effects of density and season (Quinn 1988). This is equivalent to H_0(AB): $(\alpha\beta)_{ij} = 0$, i.e. no interaction between factor A and factor B; the effect of A is the same at all levels of B and the effect of B is the same at all levels of A. For example, the effect of density on the mean number of egg masses per limpet is the same in both seasons and the effect of season on the mean number of egg masses per limpet is the same for all four densities.

F-RATIOS

We can test these H_0s by seeing which of our mean squares have the same expected value when the H_0 is true (Table 9.9). The *F*-ratios for testing these H_0s are provided in Table 9.11. It is clear that MS_A and $MS_{Residual}$ have the same expected value when there is no effect of factor A so these two mean squares are used in an *F*-ratio to test the H_0(A). MS_B and $MS_{Residual}$ have the same expected value when there is no effect of factor B so these two mean squares are used in an *F*-ratio to test the H_0(B). Finally, MS_{AB} and $MS_{Residual}$ have the same expected value when there is no effect of the interaction between A and B so these two mean squares are used in an *F*-ratio to test the H_0(AB).

The degrees of freedom associated with these *F*-ratios are simply the df associated with the two terms. For example, the df for the *F*-ratio testing the interaction H_0 are $(p-1)(q-1)$ and $pq(n-1)$. The *F*-ratios are compared to an *F* distribution and conclusions about whether to reject or not reject

Table 9.11 | *F*-ratios used for testing main effects and interactions in a two factor ANOVA model for different combinations of fixed and random factors

Source	A and B fixed	A and B random	A fixed, B random Restricted version	A fixed, B random Unrestricted version
A	$\frac{MS_A}{MS_{Residual}}$	$\frac{MS_A}{MS_{AB}}$	$\frac{MS_A}{MS_{AB}}$	$\frac{MS_A}{MS_{AB}}$
B	$\frac{MS_B}{MS_{Residual}}$	$\frac{MS_B}{MS_{AB}}$	$\frac{MS_B}{MS_{Residual}}$	$\frac{MS_B}{MS_{AB}}$
AB	$\frac{MS_{AB}}{MS_{Residual}}$	$\frac{MS_{AB}}{MS_{Residual}}$	$\frac{MS_{AB}}{MS_{Residual}}$	$\frac{MS_{AB}}{MS_{Residual}}$

Box 9.8 | Terminology used for identifying fixed and random effects in expected mean squares

- D_p, D_q and D_r reflect the terminology presented in Winer *et al.* (1991) for fixed and random factors. $D_p = 1 - p/P$, where p is the number of levels of factor A, and P is the possible number of levels. q and r denote the levels of factors B and C, respectively. If A is a fixed factor, then the p levels represent all possible levels, so $p = P$ and $D_p = 0$. If A is random, the p levels are assumed to be a (very small) sample of a population of possible levels, and $p/P = 0$, so $D_p = 1$.
- n represents the number of replicates at each combination of A, B and C.
- Terms associated with factors A, B and C are denoted by Greek letters α, β and γ, respectively.
- σ_α^2 refers to an added variance component when factor A is random and to the variance between fixed A group means ($\sum_{i=1}^{p} \alpha_i^2/(p-1)$) when factor A is fixed. Similar definitions apply for other terms.

the H_0 are drawn in the usual manner. The worked example from Quinn (1988) in Box 9.4 illustrates these tests for a fixed effects model.

Random effects models

With random factors, our focus is on tests of added variance components, rather than differences between the means of the chosen groups.

FACTOR A

H_0(A): $\sigma_\alpha^2 = 0$, i.e. no added variance due to all possible levels of factor A that could have been used. For example, there is no added variance in the number of meals eaten by sawfly larvae due to all possible species of sawflies that Kause *et al.* (1999) could have used.

FACTOR B

H_0(B): $\sigma_\beta^2 = 0$, i.e. no added variance due to all possible levels of factor B that could have been used. For example, there is no added variance in the number of meals eaten by sawfly larvae due to all possible trees that Kause *et al.* (1999) could have used.

INTERACTION BETWEEN A AND B

H_0(AB): $\sigma_{\alpha\beta}^2 = 0$, i.e. no added variance due to any of the interaction effects between all possible levels of factor A and factor B that could have been used. For example, there is no added variance in the number of meals eaten by sawfly larvae due to any interaction between all possible species of sawflies and all possible trees that Kause *et al.* (1999) could have used in their study.

F-RATIOS

We can again test these H_0s by seeing which of our mean squares have the same expected value when the H_0 is true (Table 9.9, Table 9.11). The *F*-ratio for H_0(A) uses MS_A and MS_{AB}, because the expected value for MS_A includes the interaction variance. The *F*-ratio for H_0(B) uses MS_B and MS_{AB}, because the expected value for MS_B also includes the interaction variance. The *F*-ratio for H_0(AB) uses MS_{AB} and $MS_{Residual}$ as in the fixed effects model.

Mixed effects models

The null hypotheses for main effects in a mixed model are basically the same as those in fixed and random effects models, for fixed and random factors respectively. Let us assume that factor A is fixed and factor B is random and that we are using the traditional Model I values of the expected mean squares, i.e. imposing constraints on the interaction terms (Box 9.7).

FACTOR A (FIXED)

H_0: $\mu_1 = \mu_2 = \ldots = \mu_i = \ldots = \mu_p$. This H_0 states there is no difference between the marginal means for Factor A pooling over the levels of factor B. For example, no difference in the mean density of oysters per quadrat for each zone, pooling over all possible randomly chosen sites (Minchinton & Ross 1999). This is equivalent to H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_i = 0$, i.e. no effect of any level of factor A pooling over the levels of factor B. For example, no effect of any of the four zones on mean density of oysters per quadrat, pooling all possible sites.

FACTOR B (RANDOM)

H_0: $\sigma_\beta^2 = 0$, i.e. no added variance due to all possible levels of factor B that could have been used. For example, there is no added variance in the density of oysters per quadrat due to all possible sites that Minchinton & Ross (1999) could have used, pooling the four zones.

INTERACTION BETWEEN A AND B

The null hypothesis for the interaction term, which is considered a random variable even though it is an interaction between a fixed effect and a random variable, is H_0: $\sigma_{\alpha\beta}^2 = 0$, i.e. no added variance due to any of the interaction effects between the fixed levels of factor A and all possible levels of factor B that could have been used. When either factor is random, then the interaction is random because it represents a subset (depending on the levels of the random factor chosen) of all the possible interactions (Underwood 1997). For the Minchinton & Ross (1999) study, this H_0 is that there is no added variance in the density of oysters per quadrat due to any interaction between the fixed zones and all possible sites that could have been used.

F-RATIOS

We again test these H_0s by seeing which of our mean squares have the same expected value when the H_0 is true (Table 9.10). The *F*-ratios for

Table 9.12 | Main effects and interaction effects as sets of contrasts among marginal and cell means

(a)

	B_1	B_2	B_3	A marginal means
A_1	μ_{11}	μ_{12}	μ_{13}	μ_{A1}
A_2	μ_{21}	μ_{22}	μ_{23}	μ_{A2}
A_3	μ_{31}	μ_{32}	μ_{33}	μ_{A3}
B marginal means	μ_{B1}	μ_{B2}	μ_{B3}	

(b) H_0: no effects of A

Set 1	Set 2	Set 3
H_0: $\mu_{A1} - \mu_{A2} = 0$	H_0: $\mu_{A1} - \mu_{A2} = 0$	H_0: $\mu_{A1} - \mu_{A3} = 0$
H_0: $\mu_{A2} - \mu_{A3} = 0$	H_0: $\mu_{A1} - \mu_{A3} = 0$	H_0: $\mu_{A2} - \mu_{A3} = 0$

(c) H_0: no interaction effects

Effect of A same at each level of B		Effect of B same at each level of A	
Set 1	Set 2	Set 1	Set 2
$\mu_{11} - \mu_{21} - \mu_{12} + \mu_{22} = 0$	$\mu_{11} - \mu_{21} - \mu_{12} + \mu_{22} = 0$	$\mu_{11} - \mu_{12} - \mu_{21} + \mu_{22} = 0$	$\mu_{11} - \mu_{12} - \mu_{21} + \mu_{22} = 0$
$\mu_{12} - \mu_{22} - \mu_{13} + \mu_{23} = 0$	$\mu_{11} - \mu_{21} - \mu_{13} + \mu_{23} = 0$	$\mu_{21} - \mu_{22} - \mu_{31} + \mu_{32} = 0$	$\mu_{11} - \mu_{12} - \mu_{31} + \mu_{32} = 0$
$\mu_{21} - \mu_{31} - \mu_{22} + \mu_{32} = 0$	$\mu_{11} - \mu_{31} - \mu_{12} + \mu_{32} = 0$	$\mu_{12} - \mu_{13} - \mu_{22} + \mu_{23} = 0$	$\mu_{11} - \mu_{13} - \mu_{21} + \mu_{23} = 0$
$\mu_{22} - \mu_{32} - \mu_{23} + \mu_{33} = 0$	$\mu_{11} - \mu_{31} - \mu_{13} + \mu_{33} = 0$	$\mu_{22} - \mu_{23} - \mu_{32} + \mu_{33} = 0$	$\mu_{11} - \mu_{13} - \mu_{31} + \mu_{33} = 0$

testing these H_0s are provided in Table 9.11. It is clear that MS_A and MS_{AB} have the same expected value when there is no effect of factor A so these two mean squares are used in an *F*-ratio to test the H_0(A). In contrast, the *F*-ratio for the H_0(B) of no effect of the random factor B uses MS_B and $MS_{Residual}$. Finally, the *F*-ratio for the H_0(AB) of no effect of the interaction between A and B uses MS_{AB} and $MS_{Residual}$ as in the fixed and random effects models.

9.2.4 What are main effects and interactions really measuring?

Fixed effects models

Main effects and interactions can be considered as a set of orthogonal (independent) contrasts between marginal means or cell means (Table 9.12(a)). The SS_A is simply the sum of the SS for two independent contrasts among the A marginal means. With three levels of A, the two contrasts in any of the three sets in Table 9.12(b) make up the main effect of A. Similar contrasts can be generated for factor B. Remember from Chapter 8 that the number of independent contrasts possible will be the number of df for that factor. Contrasts between cell means can be determined for interaction effects. For example, if we consider the interaction as the effect of A at each level of B, then one set of four independent contrasts would include the difference between A_1 and A_2 at B_1 and B_2 and at B_2 and B_3, and the difference between A_2 and A_3 at B_1 and B_2 and at B_2 and B_3, as indicated in the first column of Table 9.12(c). The sum of the SS for these contrasts will be SS_{AB}. There are other sets of independent interaction contrasts, for both the effect of A at each level of B and the effect of B at each level of A. The sum of the SS for the contrasts within any of these sets will be SS_{AB}. Milliken & Johnson (1984) provide clear examples and formulae for determining these contrasts. Understanding these contrasts is important when

Table 9.13 Illustration of interactions for an artifical two factor design with two levels of each factor.

(a)

	B_1	B_2	Marginal A means	$\beta\ (\mu_2 - \mu_1)$
A_1	5	12.5	8.75	7.5
A_2	10	17.5	13.75	7.5
Marginal B means	7.5	15	11.25	
$\alpha\ (\mu_2 - \mu_1)$	5	5		

(b)

	B_1	B_2	Marginal A means	$\beta\ (\mu_2 - \mu_1)$	$\beta\ (\log\mu_2 - \log\mu_1)$
A_1	5 (0.699)	10 (1.000)	7.5	5	0.301
A_2	10 (1.000)	20 (1.301)	15	10	0.301
Marginal B means	7.5	15	11.25		
$\alpha\ (\mu_2 - \mu_1)$	5	10			
$\alpha\ (\log\mu_2 - \log\mu_1)$	0.301	0.301			

(c)

	B_1	B_2	Marginal A means	$\beta\ (\mu_2 - \mu_1)$	$\beta\ (\log\mu_2 - \log\mu_1)$
A_1	5 (0.699)	20 (1.301)	12.5	15	0.602
A_2	10 (1.000)	10 (1.000)	10	0	0.000
Marginal B means	7.5	15	11.25		
$\alpha\ (\mu_2 - \mu_1)$	5	−10			
$\alpha\ (\log\mu_2 - \log\mu_1)$	0.301	−0.301			

Note:
Middle entries are cell means with $\log_{10}$ values in parentheses. Marginal means are also provided. (a) No interaction (all $(\alpha\beta)_{ij} = 0$) where the effect of A is the same at each level of B and vice versa. (b) Simple interaction (all $(\alpha\beta)_{ij} = \pm 1.25$) where the effect of A is greater at B_2 compared to B_1 and the effect of B is greater at A_2 compared to A_1. Note interaction effects removed by log transformation (all $(\alpha\beta)_{ij} = 0$). (c) More complex interaction (all $(\alpha\beta)_{ij} = \pm 3.75$) where the effect of A is in the opposite direction at B_2 compared to B_1 and there is no effect of B at A_2 but a strong effect at A_1. Note interaction not removed by log transformation.

we are dealing with designs with missing cells (Section 9.2.6). Note that with unbalanced designs (unequal sample sizes), the sum of the contrasts won't add to the SS for the relevant factor or interaction (Section 9.2.6).

Let's look at the meaning of the interaction term in more detail, by comparing three possible configurations of cell means in a design with two levels of two fixed factors (Table 9.13). In the first example (Table 9.13(a)), there are effects of A and B (both sets of marginal means differ) but no interaction between the two factors. The effect of A is the same at each level of B (a change by five units) and the effect of B is the same at each level of A (a change by 7.5 units). All the $(\mu_{ij} - \mu_i - \mu_j + \mu)$ equal zero, indicating no interaction. No interaction indicates the effects of factor A and factor B are additive and independent of each other, i.e. the response variable can be predicted by just the two main effects.

In the second example (Table 9.13(b)), there are also effects of both A and B as both sets of marginal means differ. More importantly, there is an interaction between the two factors. The effect of A is different at each level of B (five unit change at B_1 and ten unit change at B_2) and the effect of B is different at each level of A, although the effects are consistently in the same direction. The differences between marginal means do not simply reflect the effects of each factor. All the $(\mu_{ij} - \mu_i - \mu_j + \mu)$ equal ± 1.25, indicating an interaction.

In the third example (Table 9.13(c)), there is a more complex interaction. Note that the marginal means indicate only a small effect of A (minus 2.5 units from A_1 to A_2). It is clear from the cell means, however, that there is actually an opposite effect of A at each level of B (plus five unit change at B_1 and minus ten unit change at B_2). A similar result occurs for B. The marginal means suggest a strong effect (plus 7.5 units from B_1 to B_2), whereas the cell means show only a strong effect of B at A_1 and no effect at A_2. Neither set of marginal means represents a consistent effect for each factor. All the $(\mu_{ij} - \mu_i - \mu_j + \mu)$ equal ± 3.75, indicating a stronger interaction than the previous example. In both Table 9.13(b) and (c), the interaction indicates the effects of factor A and factor B are multiplicative, i.e. the response variable cannot be predicted by just the two main effects.

If there are interactions then interpretation of main effects becomes more difficult. Remember that an interaction is telling us that the main effects are not independent of each other, i.e. the effect of one factor depends on the levels of the

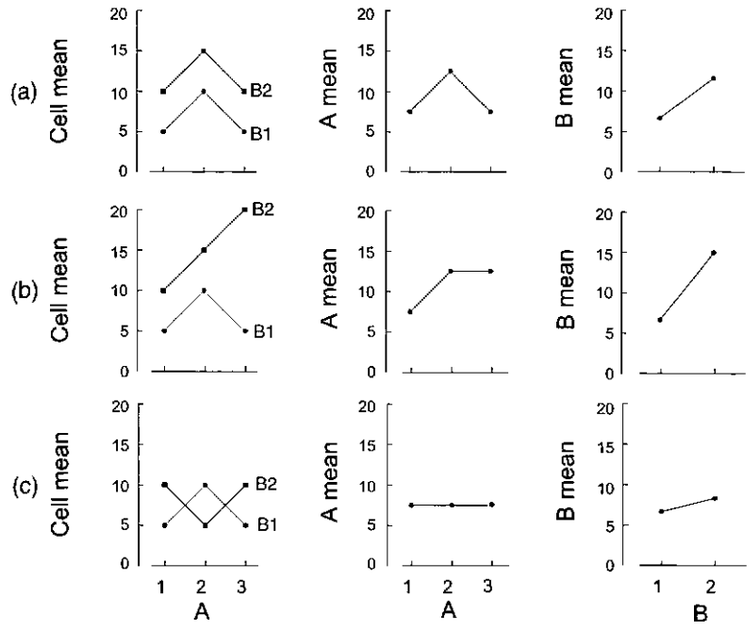

Figure 9.3 Interaction plots of the cell means of two factors, A (A1, A2, A3) and B (B1, B2) and plots of marginal main effects means for each factor separately: (a) no interaction, (b) moderate interaction, and (c) severe interaction.

second factor. The main effect of a given factor (comparison of marginal means) pools over the levels of the other factor, which is not appropriate if the effects of the two factors are not independent. Figure 9.3 shows a range of interactions between two factors; note that interactions can be moderate (B2 greater than B1 for all levels of A but relative size of difference varies) or severe (B2 greater than B1 for A1 and A3 but this difference is reversed for A2). Underwood (1981, 1997) provided clear examples of how large interactions can result in misleading interpretation of non-significant main effects in an ecological context. He showed how a strong interaction could result in non-significant main effects when the main effects were actually strong, they were just not consistent across the levels of the other factor.

This suggests that there should be a sequence of hypothesis tests for fixed effects factorial ANOVAs. Most textbooks recommend testing the H_0 of no interaction first. If this test is not significant, then tests of main effects can proceed. If the interaction is significant, then tests of main effects will be difficult to interpret. Neter *et al.* (1996) suggested a modification of this strategy. They argued that interactions can still be significant without precluding interpretation of main effects and recommended seeing if interactions are "important" before deciding whether main effects can be examined, although defining important interactions is subjective. One of the arguments for still interpreting main effects in the presence of an interaction is that it is difficult to envisage a significant interaction producing

significant main effects where no real effects exist. Indeed, the examples in Underwood (1981, 1997) illustrated misleading non-significant main effects in the presence of an interaction.

We generally agree with the traditional approach that says that main effects may be difficult to interpret in the presence of statistically significant interactions when all factors are fixed and we recommend that examining the nature of the interaction is the most sensible strategy when it is clearly significant. Significant main effects may still be of interest despite an interaction but common sense must be used. Non-significant main effects in the presence of interactions won't have much meaning. In fact, interactions are often of as much biological interest as the main effects. For example, in the study of Quinn (1988), the interaction between density and season was of most interest because it would reflect changing effects of intraspecific competition when food availability changes. Interactions should not just be treated as a nuisance in factorial ANOVA models. They presumably are of considerable interest, which is why a factorial design has been used, and they nearly always offer important biological insights. There are numerous techniques for further exploring the nature of interactions in the context of factorial ANOVAs (Section 9.2.10).

Random effects models

Because the expected values of the mean squares for factors A and B both include the variance among interaction terms, the H_0s for the main effects are actually testing for a non-zero variance component over and above any random interaction effects. Therefore, the presence of an interaction does not cause problems for interpreting tests of main effects. Nonetheless, the tests of main effects will be less powerful in the presence of an interaction because the denominator of the *F*-ratio (MS_{AB}) will increase relatively more than the numerator (MS_A or MS_B). So strong interactions between the random factors will make main effects difficult to detect.

Mixed effects models

Irrespective of which version we use for a mixed model (Box 9.7, Table 9.10), the expected mean square for the fixed factor includes the variance component for the interaction term. Therefore, the test for the fixed factor actually tests for the effects of the fixed factor *over and above* the variation due to the interaction and the residual. So when one factor is random, the test of the fixed main effect is potentially interpretable even in the presence of an interaction. This is actually the justification for being able to test main treatment effects in a simple randomized blocks design even though there is no test for a block–treatment interaction, as long as the blocks factor is random (Chapter 10). Applying the Model I version of the mixed model, the tests of the random factor and the random interaction term will both use $MS_{Residual}$ as the denominator for the *F*-ratio. If we use the alternative Model II version of the mixed model, the test for the random factor changes – see discussion in Box 9.7.

Sometimes we might have a random factor with only a few levels, e.g. for practical/logistic reasons, we can only sample two or three randomly chosen sites. This causes problems because the interaction term used to test the fixed factor will not have many degrees of freedom and the test of the fixed effect may not be very powerful. This makes sense because our ability to generalize to a population of levels of a random factor should depend on how well we have sampled this population, i.e. how many levels of the random factor we use. This also explains why, when one factor is random, it is the *F* test of the fixed factor that changes. Rather than concluding whether there is an effect of the fixed factor, pooling only over the specific levels of another fixed factor, we wish to conclude whether there is a general effect of the fixed factor, pooling over all the possible levels of a random factor. We might expect such a test to be less powerful and to use a different error term.

To illustrate, consider the study by Losos (1995) who examined the survivorship of seedlings two species of palms in a coastal tropical forest in Peru. Along two randomly located transects, she defined four different successional zones running from the beach into the forest: early-seral near the beach, mid-seral and then late-seral further into the forest, and a zone dominated by a broad-leaved monocot, *Heliconia*, that may occur anywhere along the sequence. She transplanted seedlings into five plots within each zone and

Table 9.14 Two factor mixed model ANOVA from Losos (1995), where successional zone is a fixed factor and transect is a random factor. The effect of successional zone is tested against the successional zone by transect interaction with 3 and 3 df

Source	df	MS	*F*	*P*
Successional zone	3	0.060	0.31	0.819
Transect	1	0.045	3.10	0.041
Successional zone × transect	3	0.191	13.33	<0.001
Residual	30	0.014		

transect combination. A two factor model was used to analyze survivorship, with successional zone as a fixed factor and transect as a random factor (Table 9.14). The effect of successional zone, the main effect of interest, is tested against the interaction with only three and three df. In this example, the interaction was strongly significant, indicating spatial variation in the effect of successional zone on seedling survivorship. She could not reject the H_0 of no effects of successional zone over and above any variance due to the interaction.

This emphasizes an important design principle in factorial ANOVAs. When a random factor is included, it nearly always represents replication that affects the power of the test for the fixed factor (Section 9.2.13). If we cannot include many levels of the random factor, we need to decide whether we would be better to restrict our study to a single level of the random factor (e.g. a single site) but be much more confident of whether there are fixed factor effects.

9.2.5 Comparing ANOVA models

The methods we have described in Chapter 8 and Section 9.1.5 for comparing the fit of full and reduced models to test whether a particular model parameter equals zero are just as appropriate to factorial models. For example, to test the H_0 that $(\alpha\beta)_{ij}$ equals zero, i.e. no interaction in a fixed effects model, we can compare the fit of the full model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \quad (9.13)$$

to the fit of the reduced model that omits the term specified in the H_0:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \quad (9.19)$$

Using the example from Quinn (1988), we compare the full model:

$$\begin{aligned}&(\text{no. egg masses per limpet})_{ijk} = \mu + \\ &(\text{effect of season})_i + (\text{effect of density})_j + \\ &(\text{interaction between season and density})_{ij} + \\ &\varepsilon_{ijk}\end{aligned} \quad (9.14)$$

to the reduced model:

$$\begin{aligned}&(\text{no. egg masses per limpet})_{ijk} = \mu + \\ &(\text{effect of season})_i + (\text{effect of density})_j + \varepsilon_{ijk}\end{aligned} \quad (9.20)$$

Thus, we compare the fits of additive (no interaction) and multiplicative (with interaction) models (9.13 and 9.19). The difference in fit of these two models is simply the difference in their $SS_{Residual}$. This difference can be converted to a mean square by dividing by the difference in the $df_{Residual}$. The H_0 of no difference in fit of the two models (i.e. $(\alpha\beta)_{ij}$ equals zero; no interactive effects between the two factors) can be tested with an *F* test using $MS_{Residual}$ of the full model as the denominator. With equal sample sizes per cell, this model comparison test will produce the same result as the traditional ANOVA test. With unequal sample sizes, this equality does not hold. In practice, most statistical software uses comparison of general linear models to determine SS and MS, and test hypotheses, about specific terms in ANOVA models. This approach, in contrast to the formulae in Table 9.8, generalizes to unbalanced designs and designs with more factors, including crossed and nested, and combinations of categorical and continuous variables.

9.2.6 Unbalanced designs

Studies involving categorical predictor variables should usually be designed with equal sample sizes for two main reasons. First, hypothesis tests are much more robust to the assumptions of normality and variance homogeneity (Chapter 8, Section 9.2.8) when sample sizes are equal. Second, estimation of variance components for random effects is more difficult with unequal sample sizes. However, it is common in biology to end up with unequal sample sizes, even if the

Table 9.15 Types I, II and III SS for unbalanced two factor data

(a)

Source	df	Type I	Type II	Type III
Factor A (treatment)	1	347.145	281.077	282.752
Factor B (time)	2	2 884.336	2 884.336	2 869.265
A × B	2	131.912	131.912	131.912
Residual	23	234.400	234.400	234.400

(b)

Source	df	Type I	Type II	Type III
Factor A (location)	3	17 781.391	21 804.379	49 201.833
Factor B (functional group)	1	4 442.202	4 442.202	6 919.109
A × B	3	67 782.672	67 782.672	67 782.672
Residual	49	59 088.060	59 088.060	59 088.060

Notes:
(a) From Hall *et al.* (2000), where there are six cells with sample sizes of five in all cells except one that has only four observations. (b) From Reich *et al.* (1999), where there are eight cells with sample sizes of two, two, three, three, five, six, 15, and 21. In this example, the *F*-ratio test for functional group was only significant for Type III SS ($P = 0.020$) but not for Type I or II SS ($P = 0.061$).

study was originally designed with equal numbers of observations per cell. If the inequality of sample sizes is thought to be causally related to the factors, then it is probably useful to analyze the effect of the factors on the final number of replicates in each cell as a contingency table: see Shaw & Mitchell-Olds 1993 and Chapter 14. In many cases, unequal sample sizes are caused by random loss of observations or by practical constraints limiting the number of observations in some cells but not others. Besides the robustness issue, unequal sample sizes in factorial ANOVAs means that there is no simple additive partitioning of the SS_{Total} into components due to main effects and interactions and the formulae in Table 9.8 no longer apply. Also, the determination of expected values of mean squares can be difficult, especially when there are random factors.

Unbalanced multifactor designs are sometimes termed non-orthogonal. There are two levels of sample size imbalance in factorial ANOVAs. The first is when there are observations in every cell but the numbers of observations vary. The second, and more difficult, situation is when there are one or more cells with no observations.

Unequal sample sizes

A common situation in biology is where the sample sizes are unequal but all cells in the design have at least one observation. Again, we will focus on a two factor model, with two examples. Hall *et al.* (2000) did an experiment that examined the effects of nutrients (N and P) on the macroinvertebrate assemblages colonizing small artificial habitat units submersed in a shallow subtidal region in SE Australia. These artificial habitat units were loosely rolled sheets of porous cloth. The two factors were nutrients (two levels: control and added nutrients) and time (three levels: two, four and six months after deployment) and the response variable was species richness of macroinvertebrates. Five replicate units were collected from each treatment after each time period, except that one unit was lost on collection so one cell (control after six months) had only four replicates. The response variable is log numbers of individuals of macroinvertebrates per habitat unit. The analyses of these data are in Table 9.15(a).

Reich *et al.* (1999) examined the generality of traits of leaves from different species across a range of ecosystems and geographic regions. They sampled a number of species from different functional groups (three levels: forb, shrub, tree) and from different study sites and related ecosystems (six levels: Colorado–alpine tundra, North Carolina–humid temperate forest, New Mexico–desert grassland/woodland and pinyon–juniper woodland, South Carolina–warm temperate/subtropical forest, Venezuela–tropical rain forest, Wisconsin–cold temperate forest and prairie and alkaline fen/bog). To avoid completely missing cells, we will use a subset of their data, omitting Colorado and North Carolina and only using shrubs and trees. The response variable was specific leaf surface area, there were eight cells (four sites and two functional groups) and sample sizes (number of species) ranged from two to 21 per cell. The analyses of these data are in Table 9.15(b).

There are three different ways of calculating the SS for the main effects and the interaction when cell sizes vary, termed Types I, II and III SS. They all provide the same values of SS for the residual and interaction terms. The former is simply the sum of squared deviations between each observation and the overall mean, obtained from the $SS_{Residual}$ when the full model (with both main effects and an interaction) is fitted. The latter is based on the comparison of the fit of a full model with the fit of a reduced model without the interaction term.

The real difference between the methods for calculating SS with unbalanced designs is for the main effects and relates to the way that marginal means are calculated. The most common method is to use Type III SS that are based on unweighted marginal means and therefore are not influenced by the sample size in each cell (Table 9.15). In a model comparison framework, Type III SS for each main effect are calculated from the comparison of fitting the full model to the model without the main effect of interest. For example, to determine the SS_A, we compare the fit of the full model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \quad (9.13)$$

$$(\text{no. species})_{ijk} = \mu + (\text{nutrient})_i + (\text{time})_j + (\text{nutrient} \times \text{time})_{ij} + \varepsilon_{ijk} \quad (9.21)$$

to the fit of the reduced model:

$$y_{ijk} = \mu + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \quad (9.22)$$

$$(\text{no. species})_{ijk} = \mu + (\text{time})_j + (\text{nutrient} \times \text{time})_{ij} + \varepsilon_{ijk} \quad (9.23)$$

Many authorities on linear models recommended Type III SS for unbalanced multifactor ANOVAs (e.g. Maxwell & Delaney 1990, Milliken & Johnson 1984, Searle 1993, Yandell 1997), a recommendation for ecologists supported by Shaw & Mitchell-Olds (1993). This recommendation is because tests of main effect hypotheses using Type III SS are based on unweighted means, rather than means that depend on the sample size within specific cells. Searle (1993) pointed out that Type III SS were the equivalent of his preferred SS developed using the cell means model, although we argue that the effects model is conceptually easier for biologists to understand because the traditional main effects and interaction terms are explicit.

Type I SS are determined from the improvement in fit gained by adding each term to the model in a hierarchical sequence. For example, SS_A is determined by comparing the fit of the models:

$$y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk} \quad (9.24)$$

$$(\text{no. species})_{ijk} = \mu + (\text{nutrient})_i + \varepsilon_{ijk} \quad (9.25)$$

to the models:

$$y_{ijk} = \mu + \varepsilon_{ijk} \quad (9.26)$$

$$(\text{no. species})_{ijk} = \mu + \varepsilon_{ijk} \quad (9.27)$$

The additional SS explained by models 9.24 and 9.25 is the Type I SS_A. SS_B are determined by comparing the fit of the next two models in the sequence:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \quad (9.28)$$

$$(\text{no. species})_{ijk} = \mu + (\text{nutrient})_i + (\text{time})_j + \varepsilon_{ijk} \quad (9.29)$$

versus

$$y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk} \quad (9.24)$$

$$(\text{no. species})_{ijk} = \mu + (\text{nutrient})_i + \varepsilon_{ijk} \quad (9.25)$$

It is clear from Table 9.15 that SS_A are quite different for Type I and Type III methods, not surprising given the different pairs of models being compared. Unfortunately, there are two downsides of Type I SS. First is that the order of terms is important. The SS due to factor B will be different if it enters the model after factor A compared with

before factor A. Second, Type I SS use marginal means weighted by sample sizes and hence test hypotheses weighted by sample sizes. Most biologists would probably prefer their hypotheses to be independent of the cell sample sizes.

Type II SS are also developed from sequential model fitting. Now, however, the contribution of each term is assessed by comparing a model with that term to a model without it, but including all other terms at the same or lower level. To determine the SS_A, we compare the fit of:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \quad (9.28)$$

$$(\text{no. species})_{ijk} = \mu + (\text{treatment})_i + (\text{time})_j + \varepsilon_{ijk} \quad (9.29)$$

to

$$y_{ijk} = \mu + \beta_j + \varepsilon_{ijk} \quad (9.30)$$

$$(\text{no. species})_{ijk} = \mu + (\text{time})_j + \varepsilon_{ijk} \quad (9.31)$$

The additional SS explained by models 9.28 and 9.29 is the Type II SS_A. Type II SS are not dependent on the order of terms in the model but still test hypotheses of marginal means weighted by cell sample sizes. The main difference between Type II and Type III SS is that the Type III model comparison for main effects includes the interaction terms whereas the Type II model comparison doesn't.

In the example from Hall *et al.* (2000), the imbalance in the design does not affect conclusions from the *F* tests – both main effects and the interaction are significant for all three types of SS and model comparisons. However, the degree of imbalance is minor in this case. The different SS used to analyze the data from Reich *et al.* (1999) resulted in different conclusions for the hypothesis test of the main effect of location, where the *P* value based on Type III SS was 0.020 compared to 0.061 for Type I and II SS.

We prefer Type III SS for unbalanced (but not missing cells) designs, but this is still an issue of considerable debate in the statistical and ecological literature. For example, Nelder & Lane (1995) argued strongly in favor of Type I SS and recommended that hypothesis testing in linear models be based on a hierarchical series of models. They saw no role for Type III SS at all. Stewart-Oaten (1995) proposed that Type III SS are only useful for testing interactions. He argued that if interactions are absent, then test main effects with Type II SS and if interactions are present, then main effects would not be tested anyway. However, Maxwell & Delaney (1990) pointed out that this approach depends on the power of our initial test for an interaction and suggested that Type III SS are more broadly applicable.

It is very important to remember that this whole debate becomes irrelevant when factorial designs are balanced because Type I, II and III SS are identical. Therefore, we agree strongly with Underwood (1997) that unequal sample sizes should be avoided, at least by design. As Underwood (1997) pointed out, there is unlikely to be a logical reason for estimating different cell means or marginal means with different levels of precision and unequal sample sizes make variance component estimation very difficult. However, as discussed in Chapters 4 and 8, we don't recommend deleting observations to make cell sizes equal. This will reduce power, which is rarely adequate in biological experiments anyway, and the model comparison approach can easily deal with unequal sample sizes as long as we are aware of which hypotheses the different approaches are testing and we are careful about checking the assumptions of the analysis.

Missing cells

The extreme form of unequal sample sizes is where there are no observations for one or more of the cells in a multifactor ANOVA. Such data are very difficult to analyze because not all marginal and cell means can be estimated and therefore not all main effects and interactions can be tested. Type III SS based on the effects models and unweighted marginal means for main effects are inappropriate in this situation. There is no single correct analysis for missing cells designs and different approaches test different hypotheses, all of which might be of interest. The basic approach is to consider tests of main effects and interactions as sets of contrasts between marginal means and cell means respectively (Section 9.2.2). We will use two examples to illustrate analyses of factorial designs with missing cells. The first is a modification of the data from Hall *et al.* (2000) we used above, where instead of having a single habitat unit missing from the control treatment after six months, we have lost all of the observations from that cell (Table 9.16(a)). The second example is from Reynolds *et al.* (1997), who studied competition between three species of grassland plants. They identified patches dominated by each of the species, cleared a plot in each patch and seeded it with either the original species or one of the other species. Not all species–patch combinations were possible (Table 9.17(a)), so this was a design with three missing cells out of the nine possible combinations. The response variables were percentages of soil water, shoot $\delta^{13}C$ and nitrate accumulated on ion-exchange resin bags.

The best strategy is to fit a cell means model (Box 9.6), basically a single factor model for all the cells, and then test relevant contrasts based on

Table 9.16 | Analyses of two factor ANOVA design modified from Hall *et al.* (2000) with a single missing cell. (a) Design structure, (b) cell means model with tests for main effects and interactions, with contrasts for the effect of time, (c) Type III SS test of interaction of treatment by time, and (d) subset analyses, omitting 6 months and omitting controls

(a)

	2 months	4 months	6 months
Control	μ_{C2}	μ_{C4}	
Added nutrients	μ_{N2}	μ_{N4}	μ_{N6}

(b)

Source	SS	df	MS	*F*	*P*
Cells	32.013	4	8.003	93.17	<0.001
Treatment:					
Control vs nutrient added for (2 and 4) months ($\mu_{C2}+\mu_{C4}=\mu_{N2}+\mu_{N4}$)	1.063	1	1.063	12.38	0.002
Time:	34.991	3	11.664	135.62	<0.001
2 vs 4 marginal means ($\mu_{C2}+\mu_{N2}=\mu_{C4}+\mu_{N4}$)	13.441	1	13.441	156.47	<0.001
2 vs 6 for nutrient added ($\mu_{N2}=\mu_{N6}$)	19.720	1	19.720	229.57	<0.001
4 vs 6 for nutrient added ($\mu_{N4}=\mu_{N6}$)	1.830	1	1.830	21.30	<0.001
Treatment × time:					
Control vs nutrient added at 2 months vs 4 months ($\mu_{C2}-\mu_{N2}-\mu_{C4}+\mu_{N4}$)	0.491	1	0.491	5.72	0.027
Residual	1.718	20	0.086		

(c)

	SS_{Full}	df_{Full}	$SS_{Reduced}$	$df_{Reduced}$	Difference	$F_{1,20}$	*P*
Model	32.013	4	31.522	3	0.491	5.71	0.027
Residual	1.718	20	2.209	21			

(d)

	Omitting 6 months				Omitting controls			
	df	MS	*F*	*P*	df	MS	*F*	*P*
Treatment	1	1.063	12.53	0.003				
Time	1	13.441	158.34	<0.001	2	10.362	118.39	<0.001
Treatment × time	1	0.491	5.79	0.029				
Residual	16	0.085			12	0.088		

Table 9.17 Missing cells design and analysis from Reynolds *et al.* (1997). Seeds of each of the three species were planted into patches already dominated by one of the species. Only six of the nine species–patch combinations were possible. (a) Design structure, (b) ANOVA with tests of possible contrasts

(a)	Patch		
Species	1: *Plantago*	2: *Lasthenia*	3: *Calycadenia*
1: *Plantago*	μ_{11}	μ_{12}	μ_{13}
2: *Lasthenia*	μ_{21}	μ_{22}	
3: *Calycadenia*			μ_{33}

(b) Source	df
Species	2
$\mu_{13} = \mu_{33}$	1
$\mu_{11} + \mu_{12} = \mu_{21} + \mu_{22}$	1
Patch	2
$\mu_{11} = \mu_{13}$	1
$\mu_{12} + \mu_{22} = \mu_{11} + \mu_{21}$	1
Species × patch	1
$\mu_{11} - \mu_{12} - \mu_{21} + \mu_{22}$	1

cell means. The residual from that means model is used for all subsequent tests when the factors are fixed. To test the interaction effects, we need to determine which interaction contrasts are estimable, where an estimable contrast is one that doesn't rely on the missing cells. For the modified artificial habitat data from Hall *et al.* (2000), there is only one interaction contrast that is estimable ($\mu_{C2} - \mu_{N2} - \mu_{C4} + \mu_{N4}$), the effect of nutrients at two and four months, which was significant (Table 9.16(b)). There was also only one estimable interaction contrast in the plant competition data (Table 9.17(b)). If there are more than two levels of both factors, and depending on the pattern of missing cells, there may be more than one estimable interaction contrast and sums of non-estimable contrasts might also be estimable (Searle 1993). We could also use Type III SS to compare the fit of the full effects model to the fit of the reduced model; from Hall *et al.* (2000), this is the comparison of models 9.13 and 9.21 versus models 9.28 and 9.29. This *F*-ratio is testing the H_0 that all the estimable interaction contrasts are zero, and since there is only one estimable interaction contrast, this test is the same as obtained from the contrast as part of the cell means model. The point is that the *F* test based on Type III SS_{AB} does not test that all interactions between A and B are zero but only some subset that depends on the pattern of missing cells.

What about main effects? The recommended approach is to determine a set of contrasts of marginal means (for the part of the data set without missing cells) or cell means that test sensible hypotheses based on the available data. This is where the cell means model is very important. For the time factor in Hall *et al.*'s (2000) artificial habitat example, we can contrast two and four months using marginal means (H_0: $\mu_2 = \mu_4$) because all cells have observations. We can contrast two and six months and four and six months, but only using cell means for nutrient added treatments (Table 9.16(c)). For the plant competition example (Table 9.17(b)), Reynolds *et al.* (1997) contrasted the cell means for *Plantago* and *Calycadenia* for *Calycadenia* patches only (one df) and also measured the main effect of species from the analysis of the *Plantago* and *Lasthenia* combinations as a subset (one df). The combination of these two effects produced the final $SS_{Species}$ with two df. The SS_{Patch} was determined with a comparable set of contrasts. In all these cases, the residual from the cell means model was used as the denominator for all *F* tests.

Note that these analyses do not represent orthogonal partitioning of the SS_{Total}, because these designs are extreme examples of imbalance and we have already pointed out that there is no simple partitioning of the SS in unbalanced designs. We should also mention Type IV SS that are produced by the SAS statistical software package. When there are no missing cells, Type IV SS are the same as Type III SS. When there are missing cells, Type IV SS are calculated for all the estimable contrasts as described above, although SAS selects a subset of these (see Milliken & Johnson 1984 for details). Searle (1993) and Yandell (1997) have argued strongly that the default Type IV SS may not be useful for many missing cells designs. It is clearly a more sensible strategy to carefully think about what subset of hypotheses are of most interest from all those that can be tested when there are missing cells.

Another approach to missing cells designs is to analyze subsets of the data set with observations in all cells (Table 9.16(c)). In the artificial habitat example, we could delete all data from six months and fit a two factor model to the remaining data for two levels of time (two and four months) and two treatments (control and nutrients added). The SS_{AB} for this subset analysis is the same as the Type III SS from the full data set because the only estimable interaction contrast is the one from this subset (Table 9.16(c)). The first contrast of the time effect in the cell means analysis is also the main effect of time from an analysis of the subset of the data omitting six months. The other subset with observations in all cells is using added nutrient data only for all three times, although that becomes a single factor analysis. When both factors have more than two levels, and there is at least one missing cell, there may be more than one subset suitable for a factorial model. On the other hand, if there is a complex pattern of missing cells (disconnected data, *sensu* Searle 1993, Yandell 1997), then the subsets might be quite small compared to full data set, i.e. most of the rows and columns may need be to deleted to form an analyzable subset of the data.

In summary, the analysis of missing cells factorial designs will involve a combination of analyzing balanced subsets of the data, especially for interactions, and sensible contrasts of cell means for examining components of main effects. We have only discussed fixed factor models here. The tests of relevant contrasts in mixed models with missing cells are really messy because of the difficulty of calculating an appropriate error term.

9.2.7 Factor effects

When the factors are fixed (i.e. Model 1), we might wish to estimate the variance between the group means in the specific populations from which we have sampled. For factor A, this is $\Sigma_{i=1}^{p} \alpha_i^2/(p-1)$, where α_i is the difference between each group mean and the overall mean ($\mu_i - \mu$). Following Brown & Mosteller (1991), we can equate the mean squares to their expected values (the ANOVA approach) to obtain an estimate of this variance of the population group means (Table 9.18). Analogous calculations work for the other fixed effects (B and the interaction A × B) and we simply determine the proportion that each contributes to the total variance (the sum of the variance components plus the residual). An alternative for fixed factor effects is to calculate ω^2 (Hays 1994) as a measure of strength of association between the response variable and the fixed factor (see Chapter 8). This is similar to the EMS measure above except we are estimating $\Sigma_{i=1}^{p} \alpha_i^2/p$ instead of $\Sigma_{i=1}^{p} \alpha_i^2/(p-1)$, so the two measures are related by the ratio $(p-1)/p$. The formula for ω^2 given in many texts (Hays 1994, Kirk 1995) automatically determines the percentage of the total variance explained (a PEV measure, *sensu* Petraitis 1998; see Chapter 8). For fixed factor A in a two factor ANOVA:

$$\omega_A^2 = \frac{SS_A - (p-1)MS_{Residual}}{MS_{Residual} + SS_{Total}} \qquad (9.32)$$

The difference between the estimate of ω^2 and the estimate based on equating the mean squares to their expected values decreases as the number of levels of the fixed factor increases because the difference between $p-1$ and p decreases.

For a random effects model (Model 2), we wish to estimate the added variance component (the variance between the means for all possible groups) for A (σ_α^2), B (σ_β^2) and the interaction ($\sigma_{\alpha\beta}^2$). Again, we can use the ANOVA approach to estimate these variance components (Brown & Mosteller 1991, Searle *et al.* 1992) and calculate each as a proportion of the total (sum of the components plus the residual). These are equivalent calculations to those described above for fixed factors.

We emphasized in Chapter 8 that the "variance" components for fixed and random factors are interpreted differently. For a fixed factor, we are estimating the variance between group means from the specific populations we have used and the difference between the true population mean and our estimate is sampling error at the level of the replicate observations (i.e. we have used all possible groups but have sampled observations from those groups). For a random factor, we are estimating the variance between all possible

Table 9.18 | ANOVA estimates of variance components for balanced two factor ANOVA model with different combinations of fixed and random factors

A, B fixed:

Source	Expected mean square	Estimated variance component
A	$\sigma_\varepsilon^2 + nq\frac{\sum_{i=1}^{p}\alpha_i^2}{p-1}$	$\frac{MS_A - MS_{Residual}}{nq}$*
B	$\sigma_\varepsilon^2 + np\frac{\sum_{j=1}^{q}\beta_j^2}{q-1}$	$\frac{MS_B - MS_{Residual}}{np}$*
AB	$\sigma_\varepsilon^2 + n\frac{\sum_{i=1}^{p}\sum_{j=1}^{q}(\alpha\beta)_{ij}^2}{(p-1)(q-1)}$	$\frac{MS_{AB} - MS_{Residual}}{n}$*
Residual	σ_ε^2	$MS_{Residual}$

A, B random:

Source	Expected mean square	Estimated variance component
A	$\sigma_\varepsilon^2 + n\sigma_{\alpha\beta}^2 + nq\sigma_\alpha^2$	$\frac{MS_A - MS_{AB}}{nq}$
B	$\sigma_\varepsilon^2 + n\sigma_{\alpha\beta}^2 + np\sigma_\beta^2$	$\frac{MS_B - MS_{AB}}{np}$
AB	$\sigma_\varepsilon^2 + n\sigma_{\alpha\beta}^2$	$\frac{MS_{AB} - MS_{Residual}}{n}$
Residual	σ_ε^2	$MS_{Residual}$

A fixed, B random:

Source	Expected mean square	Estimated variance component
A	$\sigma_\varepsilon^2 + n\sigma_{\alpha\beta}^2 + nq\frac{\sum_{i=1}^{p}\alpha_i^2}{p-1}$	$\frac{MS_A - MS_{AB}}{nq}$*
B	$\sigma_\varepsilon^2 + np\sigma_\beta^2$	$\frac{MS_B - MS_{Residual}}{np}$
AB	$\sigma_\varepsilon^2 + n\sigma_{\alpha\beta}^2$	$\frac{MS_{AB} - MS_{Residual}}{n}$
Residual	σ_ε^2	$MS_{Residual}$

Note:
Note that the "variance component" for fixed effects (*) represents the variance between the fixed group population means.

group means and the difference between the true population variance and our estimate is sampling error at the level of groups (i.e. we have used only a sample of all the possible groups).

For mixed models, there are two approaches. The most correct method is to calculate true variance components only for the random effects in the model and determine the proportion each contributes to the total (including the residual) of the random variation in the response variable. This approach formally recognizes that the "variance" between the fixed group means is not really comparable to the added variance due to random effects. The second method (Brown & Mosteller 1991) takes a more pragmatic line and doesn't distinguish between fixed and random effects in that the equivalent of variance components are calculated for both. The argument here is that the calculations are identical for both types of effects (using the expected mean square (ANOVA) approach) and we are simply trying to apportion the total variation in the response variable amongst all the model terms in a comparable way. The different interpretations of estimates of variance between fixed group means and true added variance components still must be recognized. Any measure of proportion of explained variance for multifactor models must be treated cautiously. These measures are obviously dependent on other terms in the model (Underwood & Petraitis 1993, Underwood 1997) and are difficult to compare between analyses.

The ANOVA approach to true variance component estimation relies on equal sample sizes. When sample sizes are different, there is no straightforward solution to estimating variance components of random factors. Searle *et al.* (1992) discussed a number of modifications of the EMS (ANOVA) method, including Hendersons's Methods I, II, and III, and using cell means, but could not make definitive recommendations about which is best. They preferred maximum likelihood (ML) and restricted maximum likelihood (REML) methods of estimation for unbalanced data and these were discussed in Chapter 8. They did point out that the major limitation of these methods of estimation is their complexity and the shortage of available software.

9.2.8 Assumptions

Fortunately, the assumptions of factorial ANOVA models are basically the same as we have already discussed for single factor and multifactor nested models. The assumptions of normality and homogeneity of within-cell variances for the error terms from the model and the observations apply to hypothesis tests in factorial ANOVA models. We can check these assumptions for the observations within each cell using the same techniques (boxplots, mean vs variance plots and residuals vs cell mean plots) already described in Chapters 4 and 8.

Formal tests of homogeneity of within-cell variances, as described in Chapter 8, can be applied to factorial designs. Levene's test is probably the best and also works well when based on randomized residuals (Manly 1997). Nonetheless, our reservations outlined in Chapter 8 about using these tests in isolation from more informative diagnostic checks still hold. When the research hypotheses of interest actually concern main effects and interaction effects on variances, rather than means, modifications of the tests based on pseudo-observations (e.g. absolute residuals) described in Chapter 8 can be used (Ozaydin *et al.* 1999).

The assumption of independence is also relevant for factorial ANOVA models and the observations within each cell should be independent of each other. Problems arise if we repeatedly measure experimental or sampling units through time (see Chapters 10 and 11) or we design our experiment so that the response of some units affects the responses of others (Chapter 7).

Transformations of the response variable deserve special mention for factorial ANOVA models. Transformations of variables with skewed distributions can greatly improve normality and homogeneity of within-cell variances (Chapters 4 and 8) and should be considered when these assumptions are not met. Transformations can also affect the interpretation of interaction terms, although the effect depends on the nature of the interaction (Sokal & Rohlf 1995).

In Table 9.13(a), the effects are additive and there is no interaction between factors A and B. The difference between the two levels of A is the same at each level of B and vice versa. In Table 9.13(b), the effects of factors A and B are clearly

multiplicative and, on the raw scale, there is clearly an interaction. The difference between the two levels of A is not the same at each level of B and vice versa. However, if we log transform the cell means, this interaction effect disappears and an additive model without an interaction term would now be appropriate. After transformation, the percentage change from A_1 to A_2 is the same for both levels of B. In Table 9.13(c), the interaction is more complex, where the effects of the two levels of A are reversed for each level of B. A log transformation does not change the nature of the interaction term very much.

So a log transformation (other power transformations can also alter interaction strengths) will make effects that are multiplicative on the raw scale additive on the transformed scale (Emerson 1991, Kirk 1995, Neter *et al.* 1996, Sokal & Rohlf 1995). The decision whether to transform data before fitting multifactor ANOVA models then also depends on whether the biological interaction you are measuring is best represented on the transformed scale. An additive model after transformation is simpler but may miss multiplicative effects that represent important biological interactions. If, on the other hand, multiplicative effects are not considered biologically important interactions (i.e. only different relative percentage changes in one factor at each level of the other factor are relevant), then a log transformation to produce an additive model might be appropriate.

9.2.9 Robust factorial ANOVAs

There are few accepted robust factorial ANOVA techniques. One common approach is to use a rank transform (RT) method, whereby the data are converted to ranks and the usual ANOVA is applied to the ranks. Although this method may be useful for tests of main effects, it is inappropriate for testing interactions (McKean & Vidmar 1994, Seaman *et al.* 1994, Thompson 1991a,b) because of the nonlinear nature of rank-transformed data. The recently proposed aligned rank procedure of Salter & Fawcett (1993) may be more useful. As discussed in Chapters 3 and 8, the RT approach may not provide protection against unequal variances but can help in dealing with outliers.

The Wilcox Z test (Chapter 8) is robust to unequal variances and could be applied to factorial designs by analyzing all the cells as a one factor design, with appropriate contrasts (like a cell means model). Of course, generalized linear models (GLMs; see Chapter 13) would also be applicable when the underlying distribution of the response variable is not normal but is known to be one from the exponential family.

Randomization tests for factorial ANOVAs have been described by Edgington (1995) and Manly (1997). With both main effects and interactions involved, there are a number of different ways to randomize observations (or residuals). Observations can be randomized across all cells and either *F*-ratios or mean squares for main effects and interactions used as test statistics. Randomizing residuals across all cells and using *F*-ratios can also be used. Edgington (1995) has suggested that true randomization tests for interactions are not possible and recommended restricted randomization for testing main effects, whereby observations are randomized between groups for one factor, controlling for the other factor. Manly (1997) summarizes these and other approaches and concludes from simulations (see also Gonzalez & Manly 1998) that when based on *F*-ratios, all methods gave similar results for testing main effects and interactions, and these were similar to the classical ANOVA tests.

9.2.10 Specific comparisons on main effects

If there are no strong interactions, interpreting main effects is relatively straightforward and involves tests of marginal means, e.g. the means of factor A pooling over the levels of B and vice versa. The tests of null hypotheses of no effect of A or no effect of B can include planned contrasts and/or trend analyses or be followed by unplanned multiple comparisons, as described in Chapter 8. For example, Poulson & Platt (1996) analyzed the difference in growth between sugar maple and beech saplings (the difference was the response variable) with a two factor ANOVA model. The factors were light microenvironment (three levels: beneath canopy, single treefall gap, multiple treefall gap) and height class (three levels: small, medium, large) and they incorporated two planned contrasts for each of the main effects, although for one response variable, the interaction was significant.

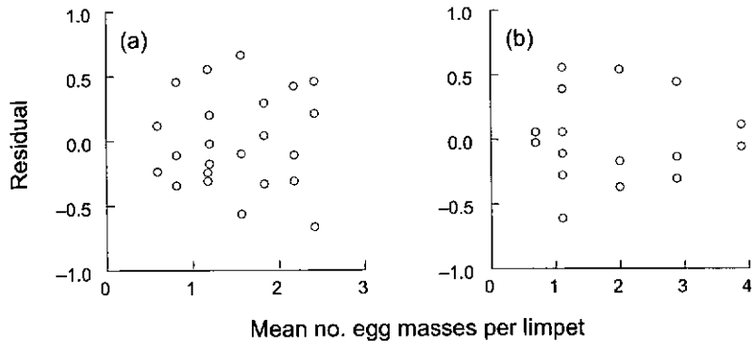

Figure 9.4 Residual plots from two factor ANOVA models for data on effects of density and season on egg mass production by *Siphonaria* limpets (Quinn 1988). (a) High shore limpets, (b) low shore limpets.

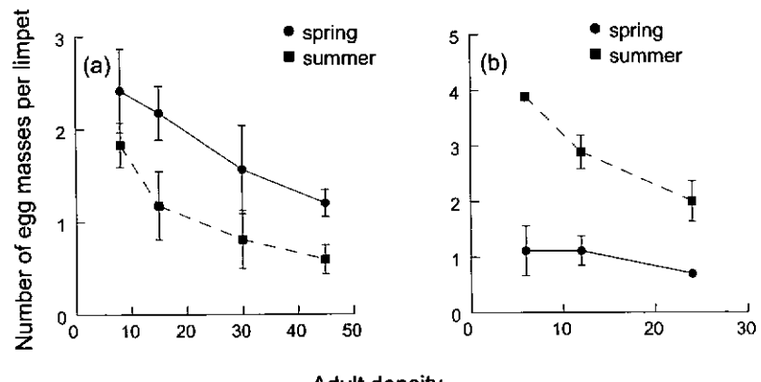

Figure 9.5 Plots of cell means and standard errors for data on effects of density and season on egg mass production by *Siphonaria* limpets (Quinn 1988). (a) High shore limpets, (b) low shore limpets.

The only tricky part of contrasts or pairwise comparisons on main effect means is to ensure that the correct error term is used if the model contains random factors. In such situations, the error term for fixed factors, and therefore any contrasts on the means of those factors, is usually an interaction term rather than the $MS_{Residual}$.

9.2.11 Interpreting interactions

The presence of an interaction between two factors is often of considerable biological interest and nearly always deserves further analysis.

Exploring interactions

Plotting cell means with the response variable on the vertical axis, the levels of one factor on the horizontal axis and lines joining the means within levels of the other factor (see Figure 9.3) is sometimes called an interaction plot. An interaction is indicated by deviation of the lines from parallel. We illustrate the effects of interactions on interpretation of main effects in Figure 9.3 (see also Table 9.13; Underwood 1997 provided a similar example and detailed explanation). In Figure 9.3(a), there is no interaction between the two factors (lines parallel) and the main effects (marginal mean plots) are straightforward to interpret. When there is a moderate interaction (Figure 9.3(b)), the marginal mean plots (and therefore tests of main effects) can become misleading. The marginal mean plot for A suggests A2 and A3 are similar whereas they are clearly different at each level of B. With complex interactions (Figure 9.3(c)), comparisons of marginal means can be completely uninterpretable. For example, the marginal mean plot for A suggests no effect when there are obvious effects at each level of B, they are just opposite.

The interaction plot for the data on the effects of adult density and season on egg production of *Siphonaria* limpets from Quinn (1988) shows no evidence of an interaction for high shore limpets but some interaction for low shore limpets (the difference between seasons is greater at density six than 12 or 24) – see Figure 9.4 and Figure 9.5. Such interaction graphs are helpful ways of understanding interactions but are necessarily subjective.

Table 9.19 (a) Final output from data sweeping of a two factor ANOVA design. Top left value is overall mean, factor A and factor B effects are in row and column borders and interaction effects are in remaining cells. (b) Example of data sweeping from Quinn (1988) showing effects of season, density and interaction on number of egg masses produced per limpet. Note that the season effects are the strongest and the interaction effects are similar to, or less than, the density effects

(a)

		B_1	B_2	etc.
	$\bar{y}$	$\bar{y}_j - \bar{y}\ (j=1)$	$\bar{y}_j - \bar{y}\ (j=2)$	
A_1	$\bar{y}_i - \bar{y}\ (i=1)$	$\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y}\ (i=1, j=1)$	$\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y}\ (i=1, j=2)$	
A_2	$\bar{y}_i - \bar{y}\ (i=2)$	$\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y}\ (i=2, j=1)$	$\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y}\ (i=2, j=2)$	
etc.				

(b)

		Density 6	Density 12	Density 24
	1.948	0.551	0.051	−0.601
Spring	−0.967	−0.413	0.087	0.324
Summer	0.976	0.413	−0.087	−0.324

Another descriptive approach is to decompose the ANOVA into a table representing the main effects and interaction effects. The general technique of splitting up the data into effects and residuals is termed sweeping (Schmid 1991) and is described for a two factor ANOVA, using the data on the effects of adult density and season on egg production of low shore *Siphonaria* limpets from Quinn (1988) as an example, in Table 9.19. The border row and column show the main effects and the central entries show the interaction effects for each cell. The season effects were stronger than the density and interaction effects, the relatively small interaction effects matching the conclusions from the ANOVA and interaction plot for these data (Box 9.4) that there is a statistically significant interaction but it does not swamp main effects.

Unplanned multiple comparison

An unplanned multiple comparison test (e.g. Tukey test) on all cell means involved in the interaction can be done (Underwood 1997). Unfortunately, there will often be many means involved and multiple comparison tests can produce ambiguous results when there are lots of groups (Chapter 8). We do not recommend this approach unless the ANOVA is exploratory and no sensible contrasts can be determined.

Simple main effects

Simple main effects test the H_0 of no effect of factor A at each level of B separately and/or no effect of factor B at each level of A separately. As an example, Stehman & Meredith (1995) described an experiment based on Radwan *et al.* (1992) who examined the effects of nitrogen (two levels: present and absent) and phosphorus (four levels: 0, 100, 300, 500 kg ha^{-1}) on growth and foliar nutrient concentrations of Douglas fir trees. This experiment would be analyzed as a two factor factorial ANOVA. Testing simple main effects might involve comparing the four P levels separately for N present and N absent or comparing N present and absent for each P level separately. The experiment examining the effect of density and season on egg mass production by limpets (Quinn 1988) showed a significant density by season interaction for low shore limpets (Box 9.4). A sensible test of simple main effects would test the density effects for each season separately.

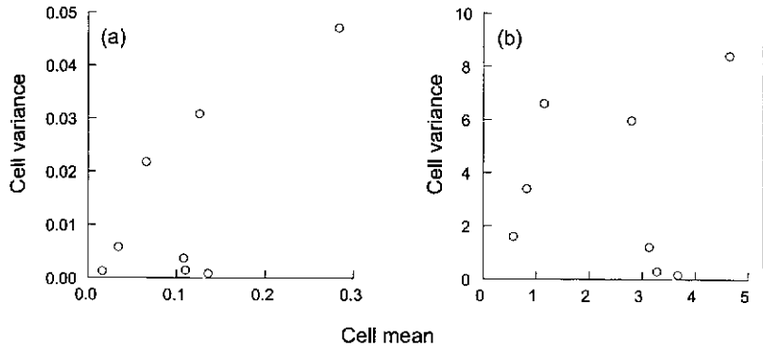

Figure 9.6 Plots of cell variances against cell means for (a) untransformed numbers of limpets per oyster and (b) square root transformed numbers of limpets per 100 oysters from Minchinton & Ross (1999). Note the weaker relationship between mean and variance after transformation.

Simple main effects don't really examine the interaction, just separate effects of one factor for each level of the other factor. In many cases, we might only wish to examine simple main effects for one of the factors. This might be particularly true if one factor is random. A significant interaction between the fixed and random factor suggests the effect of the fixed factor varies spatially or temporally and we would usually examine the simple main effects for the fixed factor at each level of the random factor separately. To illustrate from Minchinton & Ross (1999), we would test the simple main effects of intertidal zone on the density of oysters on mangrove trees for each randomly chosen site separately. When both factors are fixed, then we might want to test simple main effects for both factors. If there are many levels of one or both factors, then testing all simple main effects involves a lot of non-independent single factor ANOVAs. These are exploratory analyses looking for significant results among a collection of tests, so some correction (e.g. Bonferroni-type) to significance levels to adjust for multiple testing probably should be used (see Chapter 3).

Simple main effects tests are basically single factor ANOVAs at each level of the other factor but they are best considered as a set of particular contrasts and part of the original two factor ANOVA. The simple main effects for factor A at each level of B partition the SS and df for A and AB; simple main effects for B at each level of A partition the SS and df for B and AB. When both factors are fixed, tests of simple main effects should use the original $MS_{Residual}$ as the denominator of their F tests, because we have already decided that is our best estimate of the residual variance in our data. Simply splitting the original data and testing across the levels of A for any specific level of B with a single factor ANOVA will result in an F test with a different denominator term with different df and different power characteristics than the tests in the original two factor ANOVA. This seems inappropriate so make sure the $MS_{Residual}$ from the original factorial ANOVA is used as the error term for simple main effect tests in Model 1 ANOVAs.

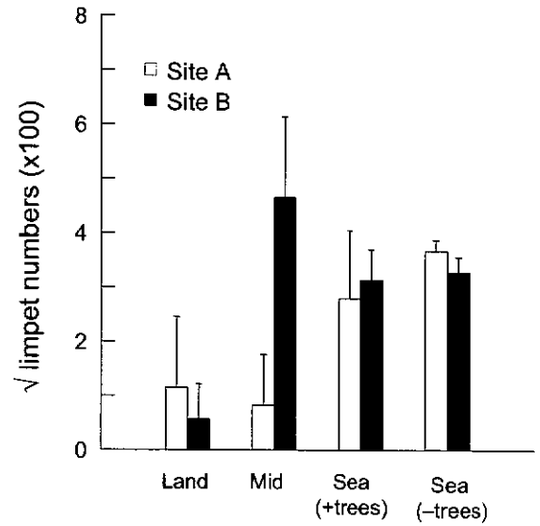

Figure 9.7 Bar graph of mean (+ standard error) square root transformed numbers of limpets per 100 oysters for different zones and sites from Minchinton & Ross (1999).

When one factor (say, B) is random (Model 3), then the main effect of A was tested against the AB interaction and the interaction was tested against the residual so there is no single denominator for tests of simple main effects of A from a partition of $(SS_A + SS_{AB})$. Under these circumstances, one approach might be to calculate a new pooled error term:

$$\frac{SS_A + SS_{AB}}{(p-1)+(p-1)(q-1)} \quad (9.33)$$

This is based on the strategy recommended by Kirk (1995) and Maxwell & Delaney (1990) for partly nested models where we have both fixed and random factors (Chapter 11). We can use this pooled error term as the denominator for simple main effect tests for factor A, although tests will only be approximate.

Keep in mind also that planned contrasts and trend analyses and unplanned comparisons can be incorporated into tests of simple main effects of fixed factors. Underwood (1997) recommended simple main effects tests to interpret interactions, although he did not use this term. He also focused on multiple comparisons, recommending SNK (or Ryan's) tests to compare the A means at each level of B separately, rather than considering the *F* tests for the simple main effects.

Treatment–contrast and contrast–contrast interactions

Treatment–contrast interactions partition the interaction term by testing contrasts (e.g. group 1 versus group 2) and trends (e.g. linear, quadratic) in one factor against the levels of the second factor. We test whether a particular contrast of groups of one factor interacts with the second factor. In the study by Stehman & Meredith (1995) examining the effects of nitrogen and phosphorus on growth and foliar nutrient concentrations of Douglas fir trees, a number of treatment–contrast interactions make sense. First, is the difference between no P (control) and the average of P_{100}, P_{300}, P_{500} consistent for treatments with N present and N absent? Second, are the linear or quadratic trends across P the same for N present and N absent? In the example for low shore limpets from Quinn (1988), we could test whether the contrast of natural density (six limpets) with increased density (12 and 24 limpets) interacts with season, i.e. whether this contrast was consistent between season (Box 9.4). Alternatively, we can test whether the linear trend in density interacts with season (Box 9.4).

Contrast–contrast interactions are a particular case of treatment–contrast interactions and test the interaction between contrasts or trends in one factor and contrast or trends in the second factor. For example, Corti *et al.* (1997) set up a factorial experiment to test the effects of hydroperiod and predation on macroinvertebrate communities in ponds on the Mississippi River floodplain. The two factors were pond (four levels: two permanently wet ponds, two temporary ponds which dried occasionally) and predator access (three levels: all access, small-fish access, no access). The design was actually slightly more complicated as there were also repeated measurements on dates (see Chapter 12) but we can just consider it a two factor analysis for the moment. They used a number of contrast–contrast interaction tests to interpret significant pond by predator interactions. For example, did the contrast between pond one and pond two (comparing the two temporary ponds) interact with the contrast of all access versus combined no access and small-fish access treatments? Did the temporary versus permanent pond contrast (ponds one and two vs three and four) interact with the contrast of all access versus combined no access and small-fish access treatments?

Both types of contrast were used by Mills & Bever (1998), who examined the effects of plant species (four levels: four species of perennial plants) and strain of pathogenic oomycete of the genus *Pythium* (six levels: control and five strains) on plant mass. Their design also included a block effect (see Chapter 10) but we can just consider the factorial component (plant species crossed with *Pythium* strain) here (Table 9.20). They included a treatment–contrast interaction test (does the effect of plant species interact with the contrast between the control and the average of the five *Pythium* strains?) and numerous contrast–contrast interaction tests (e.g. does the contrast between any two of the plant species interact with the contrast between the control and the average of the five *Pythium* strains?).

Kirk (1995) has provided computational formulae for developing such tests but they can usually be obtained from linear models routines in statistical software that allows flexible coding of contrasts.

Table 9.20 Part of ANOVA table from Mills & Bever (1998) for experiment testing effects of four plant species (An = *Anthoxanthum*, Da = *Danthonia*, Pa = *Panicum*, Pl = *Plantago*) and six pathogenic oomycete treatments (control and five strains of *Pythium*) on plant mass and root:shoot ratios

Source	df
Block	1
Plant spp.	3
Treatment	5
Contol vs average *Pythium*	1
Among *Pythium*	4
Plant spp. × Treatment	15
Plant spp. × Contol vs average *Pythium*	3
An-Da × Contol vs average *Pythium*	1
An-Pa × Contol vs average *Pythium*	1
An-Pl × Contol vs average *Pythium*	1
Da-Pa × Contol vs average *Pythium*	1
Da-Pl × Contol vs average *Pythium*	1
Pa-Pl × Contol vs average *Pythium*	1
Residual	167

Note:
Specific comparisons for treatment main effect were control treatment versus average of the five *Pythium* strains and among the *Pythium* strains. Interaction contrasts were plant species by control treatment versus average of the five *Pythium* strains ("treatment–contrast interaction") and the difference between all pairs of plant species by control treatment versus average of the five *Pythium* strains ("contrast–contrast interaction").

9.2.12 More complex designs

The two factor ANOVA model can be extended to handle more complex designs in three ways (i) three or more factor factorial designs, (ii) fractional factorial designs, and (iii) combinations of crossed and nested factors.

Complex factorial designs

Extending linear models to three or more factors is relatively straightforward, except for interpreting complex interactions. As an example, Ayres & Scriber (1994) studied climatic adaptation in caterpillars and tested the effects of sex (male, female), population (Michigan, Alaskan) and laboratory temperature (12°, 18°, 24°, 30°C) on mass of pupae produced. The three factor ANOVA model included a three factor interaction (is the interaction between temperature and population consistent for males and females?), three two factor interactions (e.g. is the difference between temperatures the same for both sexes, pooling populations?) and three main effects (e.g. is there a difference between sexes, pooling population and temperature?). Note that the three factor interaction is symmetrical – "is the interaction between temperature and population consistent for males and females?", "is the interaction between sex and population consistent across the four temperatures?", etc.

We can estimate factor effects, as either fixed factor "variances" or variance components for random factors, using modifications of the approaches described in Section 9.2.7. We can compare the relative contribution of the different main effects and interactions by equating the mean squares to their expected values as described in Section 9.2.7. Keep in mind the fundamentally different interpretation of variance components for random factors and the "variance" among fixed treatment effects. Note that when two or more random factors are included, calculation of variance components is difficult (see below).

The strategies for exploring complex interactions follow those outlined in Section 9.2.11. The equivalent of simple main effects are simple interaction effects, where the A×B interaction is examined at each level of C or the A×C interaction is examined at each level of B, etc. These simple interaction tests could then be followed by simple main effects. One difficulty is that the number of significance tests can quickly become very large when exploring complex interactions like this and some sort of Bonferroni correction to the significance levels of the tests to control the Type I error rate might be needed (Chapter 3).

If fixed and random factors are combined in these complex factorial designs, then the expected mean squares must be determined beforehand (Table 9.21) because including one or more random factors can mean that some tests will use interaction terms rather than the

Table 9.21 | Expected mean squares for three factor ANOVA model (after Winer *et al.* 1991). Factor A has *p* levels, B has *q* levels and C has *r* levels with n replicates in each cell

Source	General expected mean square
A	$\sigma_\varepsilon^2 + nD_qD_r\sigma_{\alpha\beta\gamma}^2 + nqD_r\sigma_{\alpha\gamma}^2 + nrD_q\sigma_{\alpha\beta}^2 + nqr\sigma_\alpha^2$
B	$\sigma_\varepsilon^2 + nD_pD_r\sigma_{\alpha\beta\gamma}^2 + npD_r\sigma_{\beta\gamma}^2 + nrD_p\sigma_{\alpha\beta}^2 + npr\sigma_\beta^2$
C	$\sigma_\varepsilon^2 + nD_pD_q\sigma_{\alpha\beta\gamma}^2 + npD_q\sigma_{\beta\gamma}^2 + nqD_p\sigma_{\alpha\gamma}^2 + npq\sigma_\gamma^2$
AB	$\sigma_\varepsilon^2 + nD_r\sigma_{\alpha\beta\gamma}^2 + nr\sigma_{\alpha\beta}^2$
AC	$\sigma_\varepsilon^2 + nD_q\sigma_{\alpha\beta\gamma}^2 + nq\sigma_{\alpha\gamma}^2$
BC	$\sigma_\varepsilon^2 + nD_p\sigma_{\alpha\beta\gamma}^2 + np\sigma_{\beta\gamma}^2$
ABC	$\sigma_\varepsilon^2 + n\sigma_{\alpha\beta\gamma}^2$
Residual	σ_ε^2

Note:
Coding used for expected mean squares outlined in Box 9.8.

$MS_{Residual}$ as the denominator for their *F*-ratio. This can result in reduced df and less power than anticipated for some tests. For example, as part of their study of limpets on oyster shells in mangrove forests, Minchinton & Ross (1999) used two randomly chosen sites, three zones (seaward zone with mangrove trees, middle zone with trees, and a landward zone at the upper levels) and two orientations of mangrove trunk (upper facing canopy and lower facing forest floor). There were five quadrats in each of the twelve cells and the response variable was densities of limpets per oyster surface. Although there were 48 df for the residual, the test of the interaction between the fixed factors (Zone by Orientation) used the three factor interaction with only two df as the denominator. The tests of the fixed main effects (Zone, Orientation) used the respective two factor interactions with the random factor (Zone by Site, Orientation by Site) as denominators with only two and one df respectively. To increase the power of these tests, the number of levels of the random factor (in this example, sites) needs to be increased, rather than the number of replicate observations in each cell (quadrats).

If two or three of the three factors are random (e.g. A, B and C random; see Table 9.22), then there will be no appropriate *F*-ratio tests for some terms in the model, i.e. under the H_0, there will be no other mean square with the same expected value as the term being tested. For example, in a three factor fully crossed design where all three factors are random, there are no appropriate *F*-ratios for testing for any of the main effects. There are two solutions to this problem.

1. Quasi *F*-ratios must be calculated by combining mean squares until a suitable numerator and denominator combination is found that tests the hypothesis of interest (Blackwell *et al.* 1991). For factor A in a three factor random effects model, there are two possible quasi *F*-ratios:

$$F = MS_A / (MS_{AB} + MS_{AC} - MS_{ABC}) \quad (9.34)$$

$$F = (MS_A + MS_{ABC}) / (MS_{AB} + MS_{AC}) \quad (9.35)$$

The second of these is more useful, as the first method can lead to negative *F*-ratios, which should not, by definition, occur. The degrees of freedom are also complex, and formulae are provided by Winer *et al.* (1991).

2. Alternatively, if we are primarily interested in the random factors, we can calculate confidence intervals for the variance components and see if those confidence intervals include zero (Burdick 1994).

Table 9.22 | Expected mean squares (EMS) and denominator for F-ratio test of H_0 that effect of each term equals zero for three factor crossed ANOVA model. Factor A has *p* levels, B has *q* levels, C has *r* levels with n replicates per cell. (a) Model 1 (all factors fixed) and Model 2 (all factors random). (b) One possible Model 3 (factors A and B fixed, factor C random) illustrated with example from Minchinton & Ross (1999) – see Section 9.2.12 for details

(a)

	A, B, C fixed		A, B, C random	
Source	EMS	Denominator	EMS	Denominator
A	$\sigma_\varepsilon^2 + nqr\sigma_\alpha^2$	$MS_{Residual}$	$\sigma_\varepsilon^2 + n\sigma_{\alpha\beta\gamma}^2 + nq\sigma_{\alpha\gamma}^2 + nr\sigma_{\alpha\beta}^2 + nqr\sigma_\alpha^2$	Quasi (?)
B	$\sigma_\varepsilon^2 + npr\sigma_\beta^2$	$MS_{Residual}$	$\sigma_\varepsilon^2 + n\sigma_{\alpha\beta\gamma}^2 + np\sigma_{\beta\gamma}^2 + nr\sigma_{\alpha\beta}^2 + npr\sigma_\beta^2$	Quasi (?)
C	$\sigma_\varepsilon^2 + npq\sigma_\gamma^2$	$MS_{Residual}$	$\sigma_\varepsilon^2 + n\sigma_{\alpha\beta\gamma}^2 + np\sigma_{\beta\gamma}^2 + nq\sigma_{\alpha\gamma}^2 + npq\sigma_\gamma^2$	Quasi (?)
AB	$\sigma_\varepsilon^2 + nr\sigma_{\alpha\beta}^2$	$MS_{Residual}$	$\sigma_\varepsilon^2 + n\sigma_{\alpha\beta\gamma}^2 + nr\sigma_{\alpha\beta}^2$	MS_{ABC}
AC	$\sigma_\varepsilon^2 + nq\sigma_{\alpha\gamma}^2$	$MS_{Residual}$	$\sigma_\varepsilon^2 + n\sigma_{\alpha\beta\gamma}^2 + nq\sigma_{\alpha\gamma}^2$	MS_{ABC}
BC	$\sigma_\varepsilon^2 + np\sigma_{\beta\gamma}^2$	$MS_{Residual}$	$\sigma_\varepsilon^2 + n\sigma_{\alpha\beta\gamma}^2 + np\sigma_{\beta\gamma}^2$	MS_{ABC}
ABC	$\sigma_\varepsilon^2 + n\sigma_{\alpha\beta\gamma}^2$	$MS_{Residual}$	$\sigma_\varepsilon^2 + n\sigma_{\alpha\beta\gamma}^2$	$MS_{Residual}$
Residual	σ_ε^2		σ_ε^2	

(b)

Source	Minchinton & Ross (1999)	EMS	Denominator
A	Zone	$\sigma_\varepsilon^2 + nq\sigma_{\alpha\gamma}^2 + nqr\sigma_\alpha^2$	MS_{AC}
B	Orientation	$\sigma_\varepsilon^2 + np\sigma_{\beta\gamma}^2 + npr\sigma_\beta^2$	MS_{BC}
C	Site	$\sigma_\varepsilon^2 + npq\sigma_\gamma^2$	$MS_{Residual}$
AB	Zone × orientation	$\sigma_\varepsilon^2 + n\sigma_{\alpha\beta\gamma}^2 + nr\sigma_{\alpha\beta}^2$	MS_{ABC}
AC	Zone × site	$\sigma_\varepsilon^2 + nq\sigma_{\alpha\gamma}^2$	$MS_{Residual}$
BC	Orientation × site	$\sigma_\varepsilon^2 + np\sigma_{\beta\gamma}^2$	$MS_{Residual}$
ABC	Zone × orientation × site	$\sigma_\varepsilon^2 + n\sigma_{\alpha\beta\gamma}^2$	$MS_{Residual}$
Residual		σ_ε^2	

Unfortunately, quasi *F*-ratios do not follow an *F* distribution under the H_0 and quasi-*F* tests are approximate at best (Burdick 1994). The problem becomes almost intractable if multiple random factors are combined with an unbalanced design. Our experience is that multifactor designs with more than one random factor are not common in biology, so we don't come across this situation often.

Fractional factorial designs

Sometimes we might wish to explore the effects of a number of factors but the number of combinations of factor levels is so large that the experiment is logistically impossible because it would require too many replicate units. Fractional factorial designs are often used in these situations, especially when we have a large number of factors, each with two levels. The terminology in the literature identifies this as a 2^p design where *p* is the number of two level factors. If we had four factors (*p* equals four), then the number of model terms for a fully factorial design would be 16 and the total number of experimental units required would be 16 times the number of replicates per cell. When much fewer experimental units are available and the main purpose of the experiment is to screen for important effects, a fractional factorial design might be used. There are two ways in which the required number of units

can be reduced. First, the design is nearly always unreplicated, so there is only one replicate unit within each of the cells used. By definition, this means that there is no estimate of the σ_ε^2 so some higher order interaction terms must be used as the residual for hypothesis tests. Second, the logical basis of these designs is the assumption that most of the important effects will be main effects or simple (e.g. two factor) interactions, and complex interactions will be relatively unimportant. The experiment is conducted using a subset of cells that allows estimation of main effects and simple interactions but confounds these with higher order interactions that are assumed to be trivial.

The combination of factor levels to be used is tricky to determine but, fortunately, most statistical software now includes experimental design modules that generate fractional factorial design structures. This software often includes methods such as Plackett–Burman and Taguchi designs, which set up fractional factorial designs in ways that try to minimize confounding of main effects and simple interactions.

A recent biological example of such a design comes from Dufour & Berland (1999), who studied the effects of a variety of different nutrients and other compounds on primary productivity in seawater collected from near atolls and from ocean sites. Part of their experiment involved eight factors (nutrients N, P, and Si; trace metals Fe, Mo and Mn; combination of B12, biotin and thiamine vitamins; ethylene diamine tetra-acetic acid EDTA) each with two levels. This is a 2^8 factorial experiment. They only had 16 experimental units (test tubes on board ship) so they used a fractional factorial design that allowed tests of main effects, five of the six two factor interactions and two of the four three factor interactions.

It is difficult to recommend these designs for routine use in biological research. We know that interactions between factors are of considerable biological importance and it is difficult to decide *a priori* in most situations which interactions are less likely than others. Possibly such designs have a role in tightly controlled laboratory experiments where previous experience suggests that higher order interactions are not important. However, the main application of these designs will continue to be in industrial settings where additivity between factor combinations is a realistic expectation. Good references include Cochran & Cox (1957), Kirk (1995) and Neter *et al.* (1996).

Mixed factorial and nested designs

Designs that combine both nested and factorial factors are common in biology. One design is where one or more factors, usually random, are nested within two or more crossed factors. For example, Twombly (1996) used a clever experiment to examine the effects of food concentration for different sibships (eggs from the same female at a given time) on the development of the freshwater copepod *Mesocyclops edax*. There were four food treatments, a fixed factor: constant high food during development, switch from high food to low food at naupliar stage three, the same switch at stage four, and also at stage five. There were 15 sibships, which represented a random sample of possible sibships. For each combination of food treatment and sibship, four replicate Petri dishes were used and there were two individual nauplii in each dish. Two response variables were recorded: age at metamorphosis and size at metamorphosis. The analyses are presented in Table 9.23 and had treatment and sibship as main effects. Because sibship was random, the food treatment effect was tested against the food treatment by sibship interaction. Dishes were nested within the combinations of treatment and sibship and this factor was the denominator for tests of sibship and the food treatment by sibship interaction. For age at metamorphosis, individual nauplii provided the residual term and the linear model was:

$$\begin{aligned}(\text{age at metamorphosis})_{ijkl} = \mu + \\ (\text{food treatment})_i + (\text{sibship})_j + \\ (\text{food treatment} \times \text{sibship})_{ij} + \\ (\text{dish within food treatment and sibship})_{k(ij)} + \\ \varepsilon_{ijkl}\end{aligned} \quad (9.36)$$

For size at metamorphosis, replicate measurements were taken on each individual nauplius so the effect of individuals nested within dishes nested within each treatment and sibship combination could also be tested against the residual term, the variation between replicate measurements. This linear model was:

Table 9.23 | ANOVA table for experiment from Twombly (1996) examining the effects of treatment (fixed factor) and sibship (random factor) on age at metamorphosis and size at metamorphosis of copepods, with randomly chosen dishes for each combination of treatment and sibship for age and randomly chosen individual copepods from each randomly chosen dish for size

Age at metamorphosis

Source	Denominator	df
Treatment	Treatment × Sibship	3, 42
Sibship	Dish (Treatment × Sibship)	14, 153
Treatment × Sibship	Dish (Treatment × Sibship)	42, 153
Dish (Treatment × Sibship)	Residual	153, 166
Residual		

Size at metamorphosis

Source	Denominator	df
Treatment	Treatment × Sibship	3, 42
Sibship	Dish (Treatment × Sibship)	14, 10
Treatment × Sibship	Dish (Treatment × Sibship)	42, 101
Dish (Treatment × Sibship)	Individual (Dish (Treatment × Sibship))	101, 141
Individual (Dish (Treatment × Sibship))	Residual	141, 698
Residual		

$$\begin{aligned}(\text{size at metamorphosis})_{ijklm} = \mu + \\ (\text{food treatment})_i + (\text{sibship})_j + \\ (\text{food treatment} \times \text{sibship})_{ij} + \\ (\text{dish within food treatment and sibship})_{k(ij)} + \\ (\text{individual within dish within food} \\ \text{treatment and sibship})_{l(k(ij))} + \varepsilon_{ijklm}\end{aligned} \quad (9.37)$$

Note that both models could be simplified to a two factor ANOVA model by simply using means for each dish as replicates within each treatment and sibship combination. We would end up with the same SS and F tests as in the factorial part of the complete analyses. Note also that individuals within each dish (and replicate measurements on each individual) simply contribute to the dish means but make no real contribution to the df for tests of main effects or their interaction. Power for the tests of sibship and the treatment by sibship interaction could only be improved by increasing the number of dishes and for the test of treatment by increasing the number of sibships.

Some designs require models with more complex mixtures of nested and crossed factors. For example, factor B might be nested within factor A but crossed with factor C. These partly nested linear models will be examined in Chapter 12.

9.2.13 Power and design in factorial ANOVA

For factorial designs, power calculations are simplest for designs in which all factors are fixed. Power for tests of main effects can be done using the principles described in the previous chapter, effectively treating each main effect as a one factor design. Power tests for interaction terms are more difficult, mainly because it is harder to specify an appopriate form of the effect size. Just as different patterns of means lead to different non-centrality parameters in one factor designs, combining two or more factors generates a large number of treatment combinations, and a great diversity of non-centrality parameters. Calculating the non-centrality parameter (and hence, power) is not difficult, but specifying exactly which pattern of means would be expected under some alternative hypothesis is far more difficult. Despite the difficulty specifying effects, the fixed effect factorial models have the advantage that

power for all effects is increased by increasing the number of replicates in each treatment combination, and any such steps that are taken to increase the power of a test on particular main effects will also improve power of tests of interactions. As for nested designs, interaction tests often have more degrees of freedom than corresponding main effects, so power may be more of a problem for tests of main effects.

We have already emphasized the increased complexity that can arise when random factors are included in factorial designs (see also Underwood 1997). Fixed factors and their interactions are often tested against interactions with random factors and the power of these tests will depend on the number of levels of the random factor. In the case of a two factor mixed model design, the power of tests of the random factor and the interaction will be improved by increasing the number of replicates within each combination, but the test of the fixed factor will not be improved much by this tactic. Extra care needs to be taken when designing studies that include random factors, and separate power calculations may need to be done for the fixed and random factors.

9.3 Pooling in multifactor designs

In multifactor ANOVAs with random factors, some main effects and interactions are not tested against the term with the greatest df (the Residual term). For example, in a two factor design with A fixed and B random, A is tested against B(A) if B is nested or against the AB interaction if B is crossed; in neither case is the Residual used for the test of A. What if B(A) or AB, which are tested against the Residual, are not statistically significant? Could we pool B(A) and the Residual, or AB and the Residual, to provide a test for A with more df and therefore more power?

Recommendations about whether to pool one or more non-significant sources of variation with the Residual in multifactor ANOVAs have been varied (Janky 2000). Most textbooks adopt a "sometime-pool" strategy where pooling under certain conditions is supported. The risk in pooling a nonsignificant result is that we may have made a Type II error, i.e. not rejected the H_0 that the source of variation equals zero when, in fact, it is false. For this reason, Underwood (1997) supported Winer *et al.* (1991) in suggesting that the test for the source of variation to be pooled with the Residual be done at α equals 0.25 to protect against a Type II error. Hays (1994) suggested an even more conservative approach with α equals 0.50, which corresponds to an *F*-ratio of about one, although he recommended using α equals 0.25 in practice. Sokal & Rohlf (1995) also used conservative αs (0.25, 0.50) in their pooling guidelines. We also recommend that, before pooling, any test of the H_0 that the pooled term is not different from the Residual should use a conservative α of at least 0.25.

Is there a potential cost to pooling? The main risk is that pooling terms that really do have different expected means squares will result in biased *F*-ratios for other terms that use this pooled error term. Using the pooled term as the denominator for subsequent *F*-ratios means those *F*-ratios may not necessarily follow an *F* distribution if H_0 is true. Also, we may have designed our experiment by carefully considering power required to detect a certain effect size and chosen our sample size accordingly (Chapter 7); if we then change our error term by pooling, our original design strategy and sample size may no longer be relevant. There is also some concern in the literature that a preliminary test to determine whether to pool or not may affect the power of any subsequent test (Hines 1996, Janky 2000, Kirk 1995).

Hines (1996) and Janky (2000) have recently reviewed strategies for pooling terms in ANOVA designs. Hines (1996) argued that pooling is only beneficial if another term in the ANOVA is significant after pooling but not before. We do not agree that pooling can only be beneficial if it changes the result of another test in the ANOVA. A particular term may still be non-significant after pooling, but because of greater df for the test, the probability of a Type II error is less for a given effect size than without pooling. Janky (2000) studied the effects of pooling various error terms in a partly nested model (see Chapter 12) and showed that the supposed power advantages of pooling were not always realized. Our view is that for designs with random factors, the power of tests of fixed factors can be improved by pooling nominal denominator terms of *F*-ratios with lower terms in the model. This is particularly true in field biology where the units of the random factor (either nested or crossed) are often expensive to obtain and our designs are restricted to only a few levels. However, we recommend a "sometime-pool" strategy based on a conservative test of the term to be pooled.

9.4 Relationship between factorial and nested designs

The sources of variation used in the partitioning of the total variability in the response variable depend on the experimental design. The partitioning of the nested designs we have just discussed can be related to the partitioning for a fully factorial design. For example, consider the comparison of a two factor nested (A, B within A) and a two factor factorial design. SS_A and $SS_{Residual}$ are the same in both analyses, whereas $SS_{B(A)}$ from the nested model equals SS_B plus SS_{AB} from the factorial model.

Similar equalities exist for more complex ANOVA models. For a three factor design, SS_A and $SS_{Residual}$ are again the same in both analyses, $SS_{B(A)}$ equals SS_B plus SS_{AB}, and $SS_{C(B(A))}$ equals SS_C plus SS_{AC} plus SS_{BC} plus SS_{ABC}. Nested and factorial ANOVAs are just different ways of partitioning the variability. These equalities allow nested ANOVAs to be done with software that only analyses factorial designs (Kirk 1995) but such equalities only hold for fully balanced designs.

9.5 General issues and hints for analysis

9.5.1 General issues

- Nested designs usually include levels of random subsampling nested within higher levels. Tests at each level are the equivalent of single factor ANOVAs using the group means from the level below as observations.
- We recommend Type III SS for unbalanced factorial designs because they are based on unweighted marginal means.
- If you have missing cells, you need to use cell means models and test a restricted set of hypotheses about main effects and interactions. These analyses are difficult and should be done in consultation with an experienced linear models statistician.
- Interactions are nearly always biologically important and can be further analyzed in a number of ways, including tests of simple main effects, treatment–contrast and contrast–contrast interaction tests, and less formally by cell mean plots and data sweeping.
- Rank-based tests for factorial designs should be avoided because they do not reliably detect interaction effects.
- Avoid fractional factorial designs as they must assume that certain complex interactions are negligible.
- Nested factors can be included as subsampling in factorial designs and the analyses are straightforward, although the random nested term will become the denominator for *F* tests of main effects and interactions.
- Pooling two terms in a multifactor design can increase the power of some tests. However, test the equality of the two terms to be pooled with a conservative significance level, e.g. 0.25.

9.5.2 Hints for analysis

- Make sure that when testing the H_0 for factor A in a nested design that you use the B(A) term for the denominator of the *F* test if B is random. Your favorite software may default to testing all terms against the residual.
- To increase power of the test for factor A in a two or more factor nested design, you need to increase the number of levels of B within each level of A. Increasing the number of levels of lower factors won't help much.
- When including random factors in factorial designs, ensure that you have worked out the expected mean squares and you know which terms are used as denominators for *F* tests of fixed factors and interactions. You may not have as many df as you think and might need to increase the number of levels of the random factor, which basically forms part of the replication in these designs.
- When testing simple main effects and treatment–contrast or contrast–contrast interaction tests, make sure you use the appropriate term as the denominator of the *F* test. When all factors are fixed, this will be the $MS_{Residual}$ from fitting the orginal factorial ANOVA model.

Chapter 10

Randomized blocks and simple repeated measures: unreplicated two factor designs

In Chapter 9, we described the analyses of completely randomized (CR) designs where the factors were either crossed with, or nested in, others. There are several other experimental designs that have special analytical requirements, and are used very commonly in the biological sciences. These include unreplicated factorial designs and designs that combine crossed (factorial) and nested arrangements. We deal with these two groups of designs in the next two chapters. In most cases, the main aim of these designs is to reduce the unexplained variation ($MS_{Residual}$) compared to a CR design. Such designs can be more efficient than CR designs, i.e., they offer more precise estimates of parameters and more powerful tests of the null hypotheses of interest, with no increase in the overall resources needed for the experiment. In contrast to CR designs, however, they involve restricted randomization of factor levels to experimental units and usually have additional assumptions. We will consider the simplest of these designs in this chapter.

We also recommend that biologists distinguish between the physical design (or structure) of an experiment and the linear model used to analyze it. The same model can be applied to a number of different experimental designs and we find some of the literature on these analyses confusing because the label used for the design is often interchanged with the label used for the analysis.

10.1 Unreplicated two factor experimental designs

A class of experimental designs commonly used in biology is based on a two factor crossed (factorial) design with a single observation in each cell. A completely randomized version of this design, where one experimental unit is allocated randomly to each combination of the two factors, and both factors are of equal interest, is rarely used in biology. This is because interactions between the two factors are likely to be of some interest in such settings and interactions cannot easily be detected without replication in each cell. Such experimental designs might only be useful in exploratory experiments where interactions are unlikely, such as industrial settings (Milliken & Johnson 1984). The linear model for a two factor crossed ANOVA with one observation per cell can, however, be used to analyze two types of experimental design that are very common in biological research – randomized complete blocks (RCB) and simple repeated measures (RM) designs. Although the physical structure of these types of experiments is different, Kirk (1995), Mead (1988) and others have emphasized, as we do in this chapter, that the appropriate null hypotheses and linear models are identical.

10.1.1 Randomized complete block (RCB) designs

These are experimental or sampling designs where one factor is a "blocking" variable and the other factor is the main treatment factor of interest. The

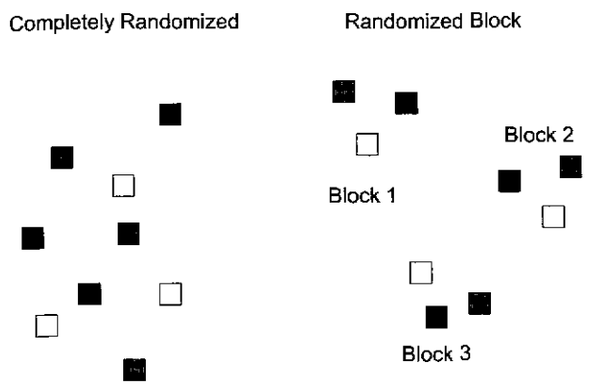

Figure 10.1 Spatial layout of an experiment with three treatments (three levels of treatment factor) and three experimental units for each treatment, contrasting a completely randomized design and a randomized blocks design.

basic principle of blocking in experimental design is to group experimental units together into blocks (which are usually units of space or time) and then each level of the treatment factor(s) is applied to one experimental unit in each block (Figure 10.1). Such designs are used when we suspect that the background environment is patchy enough to increase the variation in the response variable substantially. If experimental units are placed randomly through space (or time), we may get such high levels of background variation as to obscure any effects of the factor of interest. If we group the experimental units into "blocks" that have similar background conditions (e.g. because they are closer together in space or time), we might be able to explain some of the total variation in the response variable by differences between blocks and thus reduce the residual (unexplained) variation. This will permit more precise estimates of parameters and more powerful tests of treatments. Although blocks are commonly spatial groupings, blocks may also represent experimental units matched by physical or biological characteristics that do not have to be grouped in space or time, e.g. using organisms of similar size or age, plots of ground with similar soil characteristics.

Randomized block designs are common in the biological literature.

- Robles *et al.* (1995) examined the effect of increased mussel (*Mytilus* spp.) recruitment on seastar numbers on a rocky shore. There were two treatments: 30 – 40 l of *Mytilus* (0.5–3.5 cm long) added, no *Mytilus* added. Four matched pairs of mussel beds were chosen, each pair representing a block. Treatments were assigned randomly to mussel beds (experimental units) within a pair (block).
- Faeth (1992) applied one of four leaf damage treatments to four branches within eight randomly chosen trees of an evergreen oak. Trees were blocks and leaf damage was the treatment. Each damage treatment was represented once (on a single branch, the experimental unit) in each block.
- Evans & England (1996) applied one of three artificial honeydew treatments (honeydew followed by water ten days later, water followed by honeydew, water followed by water as control) to three plots in each of ten rows (30 plots in total) in a cultivated alfalfa field. The treatment was type of honeydew application and each row of three plots was a block, with ten blocks in total; a plot was the experimental unit. This experiment also included two repeated measurements, although we can ignore these for the purposes of this chapter and imagine analyzing either of them two times, or the difference between them, as the response variable in a RCB design.

It is apparent from these examples that blocks can be established in two ways (see also Newman *et al.* 1997). First, experimental units may be grouped into blocks at a spatial scale chosen by the investigator as part of the experimental design. The success of RCB designs then depends on establishing blocks at a scale that explains some of the variation in the response variable. Evans & England (1996) used plots (experimental units) in rows (blocks) and the spatial scale of plots within rows and between rows was determined by

the investigators. Second, experimental units may be fixed in time or space and blocks are naturally occurring groups of such units and their scale is not under control of the investigator. Faeth (1992) used branches (experimental units) on trees (blocks) and the spatial scale of neither was under the investigator's control.

In RCB designs, factor levels are randomly applied to separate experimental units within each block. This design was originally developed for agricultural experiments where blocks are often paddocks (or fields) that are subdivided for the application of factor levels. RCB designs also extend logically to split-plot experiments (Chapter 11), where another set of factor levels is applied to the whole blocks in addition to the treatments within blocks. Note that the RCB design can also be compared to an equivalent-sized single factor design (factor equals treatments) in which the residual is split into variation due to blocks, representing an attempt to control "nuisance" variables related to the scale of blocking, and the remainder. RCB designs involve a restriction on randomization, in contrast to a CR two factor design (Hicks & Turner 1999). Randomization for the RCB design is restricted to experimental units within each block whereas for a CR two factor crossed design with one observation per cell, randomized allocation of experimental units is to all combinations of the two factors, i.e. randomization across both factors.

Mites and domatia on leaves

Walter & O'Dowd (1992) were interested in testing the hypothesis that leaves of the shrub *Viburnum tinus* with domatia (small shelters at the juncture of veins on leaves) have more mites than leaves without domatia. Fourteen paired leaves (blocks) on a shrub of *V. tinus* were randomly chosen and one leaf in each pair had its domatia shaved while the other remained as a control; the number of mites was recorded on each leaf (experimental unit) after two weeks. The analyses of this experiment are in Box 10.1.

Box 10.1 Worked example of randomized complete block analysis: mites on leaves

Walter & O'Dowd (1992) examined the role of domatia (small shelters at the juncture of veins on leaves) in determining the numbers of mites on leaves of plant species with domatia. They did an experiment using 14 pairs of leaves (randomly chosen) with one leaf in each pair with shaved domatia and the other as a control (normal domatia). The response variable was total number of mites per leaf, which Walter & O'Dowd (1992) transformed to $\log_e(0.5 + (\text{mite} \times 10))$, ostensibly to improve normality and homogeneity of variances between treatments, the 0.5 added because of zeros although multiplication by ten seemed unnecessary. The data were analyzed using model 10.1, the factors being block and treatment and the response variable being $\log_e(0.5 + (\text{mite} \times 10))$.

The main H_0 of interest was that there was no effect of shaving domatia on the mean $\log_e(0.5 + (\text{mite} \times 10))$ per leaf, pooling across all possible blocks.

Source	SS	df	MS	F	P
Treatment	31.341	1	31.341	11.315	0.005
Block (leaf pair)	23.058	13	1.774	0.640	0.784
Residual	36.007	13	2.770		

The ANOVA showed that the H_0 of no effect of domatia shaving, averaging over leaf pairs, should be rejected with significantly fewer mites on leaves without domatia (Figure 10.4(a)). There were no effects of blocks although given the random blocks (leaf pairs), this test is only possible if we assume no treatment by block interaction (hence the shading in the ANOVA table). We could have achieved the same test for treatment by running a "repeated measures" analysis, with block (pair) as subject and treatment as the repeated measures factor. There would be no adjusted univariate or multivariate output because there are only two treatment levels.

We also checked for the possibility of an interaction by plotting the log-transformed number of mites for each leaf against block, separating the two treatments (an "interaction" plot: Figure 10.4(a)). The number of mites on leaves without domatia was consistently less than the number on leaves with domatia for all blocks except leaf pair 3. Tukey's test for additivity did not reveal any evidence of a strong interaction:

$$MS_{non-add} = 0.0136,\ MS_{Remainder} = 2.917,\ F_{1,18} = 0.012,\ P = 0.914.$$

Interestingly, for untransformed data, $F_{non-add(1,18)} = 41.98$, $P < 0.001$, suggesting a strong block by treatment interaction. Clearly, a log transformation improved additivity, as it often does, although the difference in strength of the interaction between transformed and untransformed data is not obvious from the interaction plots (Figure 10.4(a,b)).

The plot of residuals against comparison values from a median polish clearly shows outlying values from leaf pair number 3 at the bottom left and top right of the plot (Figure 10.5(a)). This is the leaf pair that shows the opposite pattern of treatments compared with the other leaf pairs. Note that there appear fewer points than the total number of observations (28) because some observations have identical values for both axes. The plot of residuals against predicted values from the fit of the model based on means (the standard ANOVA; Figure 10.5(b)) also shows the observations from leaf pair 3 as unusual (those with residuals near 3 and −3), although not as clearly as the median-based plot. Neither plot shows any consistent pattern indicating there is no strong interaction between block (leaf pair) and treatment.

10.1.2 Repeated measures (RM) designs

This is another experimental design based on an unreplicated two factor crossed ANOVA design where factor levels are applied to whole experimental units, called subjects, or where experimental units are recorded repeatedly through time. For example, Blake *et al.* (1994) made twice-yearly bird counts of 500 m segments from ten transects in forested areas in each of Michigan and Wisconsin. The segments in each transect were separated from each other by 50 m and were treated as the experimental units in the study – some segments were omitted (because they were logged or because they were recorded at the end of an observation period when bird numbers had declined) leaving 53 segments in Wisconsin and 51 in Michigan. Separate analyses were done for each state; segments were the subjects and time was the repeated measures factor.

In repeated measures designs, treatments are applied sequentially to the whole subject, which is the equivalent of the block in RCB designs. The RM design was originally developed for psychological and/or behavioral experiments where the block or subject was usually a person. Two different terms are sometimes used for these simple RM designs (Kirk 1995).

1. Subjects × treatments designs, in which the order of factor levels is randomized for each subject. The repeated measures factor is a set of treatments that can be ordered independently of time, e.g. a set of drugs applied to experimental animals.

2. Subjects × trials designs, in which the order of factor levels cannot be randomized. The repeated measures factor is actually time, as in the example from Blake *et al.* (1994).

There are specific difficulties associated with repeated measures experiments (Neter *et al.* 1996), especially when the factor involves experimental treatments applied by the investigator (e.g. drugs given to experimental animals). The first is the problem of carryover effects, where the effect of one treatment may be affected by the preceding treatment in the sequence. This can only be solved by ensuring that the time interval between treatments is long enough to allow recovery of the "subjects". The second problem is the order or sequence effect, where measurements early in a sequence may be different from those later in a sequence, irrespective of treatment. This problem can be alleviated by randomizing the order in which a subject receives each treatment (e.g. randomizing the order in which each animal receives each drug). In many biological experiments, especially in ecology, the factor of interest is commonly time and carryover effects are not so relevant and order or sequence effects are implicit in the hypothesis being tested, e.g. differences between weeks, seasons or years. Note the absence of carryover effects does not imply absence of correlations between successive treatments in a repeated measures sequence. Repeated observations on the same subject will always be correlated to some extent and the nature of these correlations is the main determinant of the analysis strategy for these designs (see Section 10.2).

The distinction between the structure of RCB and RM designs is important. The former allocates levels of the factor of interest (treatments) randomly to different experimental units within blocks; the latter applies the treatments successively to whole blocks, commonly termed subjects, although the order of treatments can be randomized.

Burning and frog numbers in catchments

Driscoll & Roberts (1997) examined the effects of fuel-reduction burning on the abundance of a species of frog in Western Australia. They used six drainages within a catchment, which represent the subjects. In each drainage, they had a matched burnt site and control (unburnt) site and the response variable for the experiment was the difference in the number of calling male frogs between the burnt and control site in each drainage. Note that the analysis of this study could have included the burnt and control sites as an additional factor, although we will analyze the data in the way Driscoll & Roberts (1997) did, using the burnt–control difference within each drainage at each time as the response variable. This variable was recorded three times (repeated measures factor): pre-burn (1992) and two post-burn times (1993, 1994). The analyses of these data are in Box 10.2.

Box 10.2 Worked example of simple repeated measures analysis: frogs in burnt/unburnt catchments

Driscoll & Roberts (1997) examined the effects of fuel-reduction burning on the abundance of a species of frog in Western Australia. They used six drainages within a catchment, which represent the subjects or blocks. In each drainage, they had a matched burnt site and control (unburnt) site and the response variable for the experiment was the difference in the number of calling male frogs between the burnt and control site in each drainage. This variable was recorded three times (repeated measurements) – pre-burn (1992) and two times post-burn (1993, 1994). This is a classical repeated measures (subjects by trials) design.

The main H_0 of interest was that there was no difference between years in the mean difference in the number of calling male frogs between burnt and unburnt catchments.

The results from the ANOVA are as follows.

Source	df	GG–df	HF–df	MS	*F*	*P*	GG–*P*	HF–*P*
Years	2	1.42	1.83	184.722	9.660	0.005	0.0130	0.006
Residual	10	7.12	9.15	19.122				

The following are as published in Driscoll & Roberts (1997).

Source	SS	df	MS	*F*	*P*
Year	369.44	2	184.72	9.66	0.005
Block (drainage)	955.61	5	191.12	9.99	0.001
Residual	191.22	10	19.12		

Greenhouse–Geisser epsilon = 0.712, Huynh–Feldt epsilon = 0.915. Note that the Greenhouse–Geisser epsilon estimate is more conservative than the Huynh–Feldt estimate and the former results in a more severe correction of the df and a more conservative test. Although both epsilon estimates are less than one, the conclusions from the univariate ANOVA are unchanged irrespective of whether adjusted or unadjusted df and *P* values are used. We agree with the conclusion of Driscoll & Roberts (1997), that the H_0 of no difference between years should be rejected. The test of block (drainage) is only valid if we assume no year by block interaction. This test indicates significant variation between drainages.

We also included a planned contrast of the pre-burn year versus the two post-burn years, using a separate error term just for this contrast:

$F_{1,5} = 29.72$, $P = 0.003$, indicating that the post-burn years are significantly different from the pre-burn year in the burnt–control differences in the number of calling frogs.

MANOVA results:

Pillai Trace = 0.873 with 2, 4 df, $F = 13.69$, $P = 0.016$.

Mauchly sphericity test, $W = 0.5959$, chi-square approx. = 2.0709 (2 df), $P = 0.355$; Mauchly's test does not reject the H_0 of sphericity but is sensitive to non-normality.

We used some graphical checks and Tukey's test for non-additivity to see if an interaction was present. First, an "interaction" plot where blocks are along the horizontal axis and different lines/symbols represent the different years (Figure 10.6(a)). Note there is a change in the rankings of years 2 and 3 for blocks 5 and 6 but no evidence of any strong interaction. We also plotted residuals against predicted values and residuals against comparison values for the fitted additive model based on means, i.e. the standard ANOVA (Figure 10.6(b)). There is no curvilinear pattern in the first plot and no pattern at all in the second plot, suggesting that there is no strong interaction between years and blocks. The results of Tukey's test for non-additivity (see Box 10.5) were $F_{non-add} = 0.026/21.244 = 0.001$ with 1 and 9 df, $P = 0.974$, again no evidence of an interaction.

We also tested the H_0 that there was no linear trend in burnt–unburnt differences in frog numbers through the years.

Source	df	MS	F	P
Year	1	352.083	15.122	0.012
Residual	5	23.283		

Note that the error term used is different from the $MS_{Residual}$ in the original ANOVA; this is because we used a separate error term in case sphericity was not met. There is a significant linear trend from 1992 to 1994, with the difference between the burnt and control sites changing from negative to increasing positive. A quadratic trend test is also possible (years has two df) but is difficult to justify fitting a quadratic trend through three means.

Table 10.1 Data layout for a RCB design with p levels of factor A (treatments $i = 1$ to p) and q levels of factor B (blocks $j = 1$ to q) and n equals one in each cell

	A 1	A 2	A 3	A i	Block marginal means
Block 1	y_{11}	y_{21}	y_{31}	y_{i1}	$\bar{y}_{j=1}$
Block 2	y_{12}	y_{22}	y_{32}	y_{i2}	$\bar{y}_{j=2}$
Block 3	y_{13}	y_{23}	y_{33}	y_{i3}	$\bar{y}_{j=3}$
Block j	y_{1j}	y_{2j}	y_{3j}	y_{ij}	$\bar{y}_{j}$
A marginal means	$\bar{y}_{i=1}$	$\bar{y}_{i=2}$	$\bar{y}_{i=3}$	$\bar{y}_{i}$	Overall mean $\bar{y}$

Note:
From Walter & O'Dowd (1992), treatments (factor A) are leaves with domatia and shaved domatia, blocks are leaf pairs, individual leaves are the experimental units and the response variable is number of mites per leaf. For the simple RM design from Driscoll & Roberts (1997), treatments (factor A) are year (1992, 1993, 1994), blocks (i.e. subjects) are drainages, which are also the experimental units, and the response variable is difference in number of frogs between burnt and unburnt sites.

10.2 Analyzing RCB and RM designs

10.2.1 Linear models for RCB and RM analyses

Linear effects model

Consider the RCB design from Walter & O'Dowd (1992) with factor A (domatia treatment) having $i = 1$ to p being groups ($p = 2$, shaved and unshaved domatia) and factor B (leaf pairs) having $j = 1$ to q blocks ($q = 14$ leaf pairs) – see Table 10.1 and Figure 10.2. The linear model we fit to these data is an additive effects model, in which the response variable in each cell represents an additive combination of factor A (treatments) and block effects and we assume no interaction between treatments and blocks:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad (10.1)$$

Details of this linear model, including estimation of its parameters and means, are provided in Box 10.3.

Using the example from Walter & O'Dowd (1992):

$$(\text{mite number})_{ij} = \mu + (\text{domatia treatment})_i + (\text{leaf pair})_j + \varepsilon_{ij} \quad (10.2)$$

From Driscoll & Roberts (1997):

$$(\text{burnt vs unburnt difference in frog numbers})_{ij} = \mu + (\text{year})_i + (\text{catchment})_j + \varepsilon_{ij} \quad (10.3)$$

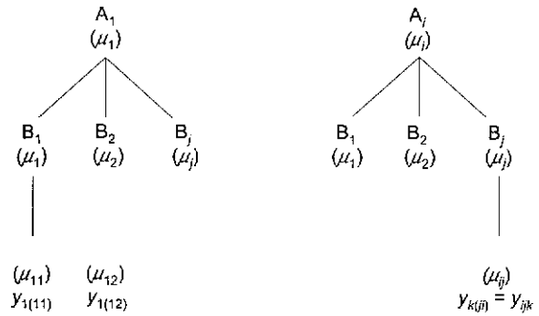

Figure 10.2 General data layout for randomized complete block ANOVA where factor A has p ($i = 1$ to p) groups and there are q ($j = 1$ to q) blocks and a single observation within each cell.

In models 10.1 and 10.2:

y_{ij} is the number of mites per leaf from the ith domatia treatment and the jth leaf pair (block).

μ is the overall (constant) mean number of mites per leaf for all combinations of domatia treatment and leaf pair (block).

If factor A is fixed, α_i is the main effect of the ith domatia treatment (removing domatia or leaving domatia) on the number of mites per leaf, pooling leaf pairs (blocks). If factor A is random, then α_i is a random variable with a variance (σ_α^2) measuring the variance in the number of mites per leaf among all possible groups that could have been used.

Box 10.3 The randomized complete block (or simple repeated measures) linear model and its parameters

Consider a RCB design with factor A ($i = 1$ to p) being treatments and factor B ($j = 1$ to q) being blocks. Each observation is y_{ij} (the value in each cell), the marginal treatment means pooling blocks are $\bar{y}_i$ and the marginal block means pooling treatments are $\bar{y}_j$ (Table 10.1). Such data structures, where we have two factors and a single observation in each cell, are sometimes referred to as two-way tables (Emerson & Hoaglin 1983). Contingency tables of frequencies (Chapter 14) are another example of a two-way table.

The linear model we usually fit to these data is an additive effects model, in which the response variable in each cell represents an additive combination of factor A (treatments) and block effects and we assume no interaction between treatments and blocks:

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

In model 10.1 we find the following.

y_{ij} is the value of the response variable from the ith level of factor A and the jth block.

μ is the (constant) overall population mean of the response variable.

If factor A is fixed, α_i is effect of ith level of factor A ($\mu_i - \mu$) pooling over blocks. If factor A 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 factor A that could have been used.

If blocks are fixed, β_j is the effect of the jth block ($\mu_j - \mu$) pooling over levels of factor A. If blocks are random, which is more common, β_j represents a random variable with a mean of zero and a variance of σ_β^2, measuring the variance in mean values of the response variable across all the possible blocks that could have been used.

ε_{ij} is random or unexplained error associated with the observation at each combination of the ith level of factor A and jth level of factor B and is

measured as $y_{ij} - \mu_i - \mu_j + \mu$. This is the error in the value of the response variable within each treatment–block combination that is not due to the treatment or block. These error terms are assumed to be normally distributed at each combination of factor A level and block, with a mean of zero [$E(\varepsilon_{ij}) = 0$] and a variance of σ_ε^2.

This model is overparameterized (see Box 8.1) so to estimate model parameters, we impose the usual restrictions that $\Sigma_{i=1}^{p} \alpha_i = 0$ if factor A is fixed and $\Sigma_{i=1}^{q} \beta_i = 0$ if blocks are fixed. Alternatively, we can fit a cell means model:

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

where μ_{ij} is the population mean for each cell and ε_{ij} is the error term associated with the observation in each cell, which we assume are normally distributed with a mean of zero and a variance of σ_ε^2. The cell means model is particularly useful when dealing with missing observations (Section 10.9).

In practice, we fit the additive effects model when analyzing RCB or simple RM designs. However, this model does not allow for an interaction between factor A (treatments) and blocks. In biological experiments, especially field experiments where blocks are spatial units, interactions between treatments and blocks are likely and we can conceptualize an alternative non-additive model that allows for an interaction between treatments and blocks:

$$y_{ij} = \mu + \alpha_i + \beta_j + [(\alpha\beta)_{ij}] + \varepsilon_{ij}$$

where μ, α_i, β_j, and ε_{ij} are defined as previously and $(\alpha\beta)_{ij}$ is the interaction between treatments and blocks. Note that the interaction term is in parentheses, because although we include it in this model, we can never estimate this term separately from the residual because we only have n equals one in each treatment–block combination. The RCB or simple RM experimental design does not permit us to separately estimate the interaction term and the error term associated with individual observations with each treatment–block combination. As Gates (1995) has pointed out, the residual or error term in a RCB design actually estimates three components: (i) block by treatment interaction, (ii) within block variability between experimental units, and potentially (iii) within experimental unit sampling variation. The important issue is that these different components cannot be distinguished because we only have one experimental unit for each treatment in each block. Although a formal test of the H_0 of no interaction is not possible, we can check for interactions in a less formal manner using graphical methods and use Tukey's test for non-additivity to detect some types of interactions (Section 10.3.2).

Conceptualizing the model in the non-additive form does have a practical use. We can include the interaction term when determining the expected mean squares for our analysis of variance and therefore assess what effect the presence of an interaction will have on the choice of F-ratios for testing both treatment and block effects.

Estimating the parameters of the factorial linear model 10.1 follows the procedures outlined for a single factor model in Chapter 8, and for nested and factorial models in Chapter 9. Consider a RCB or simple RM design, with the usual configuration of Factor A fixed and blocks/subjects random. The estimate of each cell mean μ_{ij} is simply the single observation within each cell. Estimates of the marginal means μ_i and μ_j are also straightforward. The marginal means for factor A are estimated from the observations for that level of factor A averaged across the blocks and vice versa for the marginal means for blocks. The estimate of μ is the average of all the observations, or the average of the A marginal means or the average of the B marginal means. Standard errors for these means are based on the estimate of the variance of the error terms σ_ε^2, the $MS_{Residual}$ (see Box 9.6).

The estimate of α_i is the difference between the mean of each A level and the overall mean. Interaction effects measure how much the effect of one factor depends on the level of the other factor and vice versa. If there is no interaction between the two factors, we would expect the cell means to be represented by the sum of the overall mean and the main effects:

$$\mu_{ij} = \mu_i + \mu_j - \mu$$

Therefore, the effect of the interaction between the ith level of A and jth block $(\alpha\beta)_{ij}$ can be defined as the difference between the ijth cell mean and its value we would expect if there was no interaction. This represents those effects not due to the overall mean and the main effects.

Note that in practice we don't calculate the estimated factor or interaction effects, usually focusing on contrasts of marginal or cell means (Section 10.6).

If factor B is fixed, β_j is the main effect of the specific leaf pairs (blocks) on the number of mites per leaf, pooling domatia treatments. If factor B is random, then β_j is a random variable with a variance (σ_β^2) measuring the variance in the number of mites per leaf among all possible leaf pairs (blocks) that could have been used.

ε_{ij} is random or unexplained error associated with the number of mites per leaf at each combination of the ith domatia treatment and jth leaf pair (block). This error has at least two components (Box 10.3). First, the true error due to random variability between replicate observations in the populations within each combination of treatment and block. Second, the error due to any interaction between treatment and block. With only a single observation in each block for each treatment, we cannot separately estimate these two sources of error.

Predicted values and residuals

If we replace the parameters in model 10.1 by their OLS estimates (Box 10.3), it turns out that the predicted or fitted values of the response variable from our linear model are:

$$\hat{y}_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + (\bar{y}_j - \bar{y}) = \bar{y}_i + \bar{y}_j - \bar{y} \qquad (10.4)$$

So any predicted Y-value is predicted by the marginal domatia treatment mean, the marginal leaf pair (block) mean and the overall mean. For example, the predicted number of mites per leaf for the domatia shaved treatment in leaf pair one is the marginal mean for the domatia shaved treatment (pooling leaf pairs) plus the marginal mean for leaf pair one (pooling domatia treatments) minus the overall mean number of mites per leaf.

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

$$e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_i - \bar{y}_j + \bar{y} \qquad (10.5)$$

For example, the residuals from the model relating number of mites per leaf to domatia treatment and leaf pair are the differences between the observed number of mites per leaf and the marginal mean for the domatia treatment (pooling leaf pairs) minus the marginal mean for leaf pair (pooling domatia treatments) plus the overall mean number of mites per leaf. These residuals actually estimate the effect of the interaction between blocks and treatments for each cell although this cannot be distinguished from the variation associated with each observation

Table 10.2 | ANOVA table for RCB design

Source	SS	df	MS
A (treatments)	$q\sum_{i=1}^{p}(\bar{y}_i-\bar{y})^2$	$p-1$	$\frac{SS_A}{p-1}$
B (blocks)	$p\sum_{j=1}^{q}(\bar{y}_j-\bar{y})^2$	$q-1$	$\frac{SS_B}{q-1}$
Residual	$\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{y}_{ij}-\bar{y}_i-\bar{y}_j+\bar{y})^2$	$(p-1)(q-1)$	$\frac{SS_{Residual}}{(p-1)(q-1)}$
Total	$\sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij}-\bar{y})^2$	$pq-1$	

Table 10.3 | The comparison of completely randomized and randomized block ANOVAs for the general case with p treatments and q experimental units per treatment and the example from Walter & O'Dowd (1992) with two treatments and either 14 replicates per treatment (completely randomized) or 14 blocks (block design)

	Randomized block		Completely randomized	
Source	general df	specific df	general df	specific df
Treatments	$p-1$	1	$p-1$	1
Blocks	$q-1$	13		
Residual	$(p-1)(q-1)$	13	$p(q-1)$	26
Total	$pq-1$	27	$pq-1$	27

within each cell (because n equals one for each treatment–block combination). As in all linear models, 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 10.4).

10.2.2 Analysis of variance

The classical partitioning of variation from a least squares fit of the additive effects model for a RCB or simple RM design is shown in Table 10.2. The SS are based on marginal means (Table 10.1) as for any factorial ANOVA model (see Chapter 9). SS_A measures the sum of squared differences between each treatment marginal mean and the overall mean; the SS_B measures the sum of squared differences between each block marginal mean and the overall mean; the $SS_{Residual}$ measures the sum of squared differences for a particular contrast involving cell means, marginal means and the overall mean, i.e. the interaction between treatments and blocks. The mean squares (MS) are determined by dividing the SS by their df.

The comparison between the ANOVAs for a RCB design, where experimental units are grouped into blocks, and the equivalent sized single factor CR design, where the allocation of treatments to experimental units is randomized, is shown in Table 10.3. Note that the RCB design has fewer df for the residual than the single factor CR design. The residual term in the CR design has been simply split into blocks and "residual" components. We are making a trade-off in that we are accepting fewer df in the residual term of the RCB, in expectation that the SS and MS will be lower, and more than compensate for the loss of df in terms of the power of the test of treatments (Section 10.7).

Table 10.4 | Structure of ANOVA table for "classical" repeated measures design. Note that this ANOVA is identical to a randomized blocks ANOVA, where subjects are blocks

Source	General df	Specific df
Between "subjects" (drainages)	$q-1$	5
Within "subjects" (drainages)	$q(p-1)$	12
Treatments (years)	$p-1$	2
Residual	$(q-1)(p-1)$	10
Total	$pq-1$	17

Note:
The specific example is from Driscoll & Roberts (1997) with three treatments (years) and six subjects.

For the analysis of a classical RM design, the ANOVA table is sometimes presented slightly differently compared with the analysis of a classical RCB design, to distinguish sources variation between subjects (i.e. blocks) and sources of variation within subjects – see Table 10.4. This ANOVA table is actually the same as for the usual RCB design except that within and between subjects (or blocks) sources of variation have been made explicit. The same linear model is used to analyze RCB and simple RM designs, an additive two factor ANOVA model.

The expected mean squares (EMS) for different combinations of fixed and random factors are given in Table 10.5. Note that we can derive these EMS is two ways. First, assuming there is no A by blocks interaction and fitting the standard additive model 10.1. Second, by including the possibility of an A by blocks interaction with a non-additive model (Box 10.3). In practice, we cannot really fit a non-additive model because we cannot estimate the interaction term separately from the true error. The non-additive form of the EMS, however, does allow us to evaluate the effect of an interaction on the relevant F-ratios for testing the null hypotheses.

The EMS for the non-additive model where factor A is fixed and blocks are random is based on the classical approach for mixed models (one factor fixed and one random) as outlined in Chapter 9. The interaction is considered a random effect and the interaction effects sum to zero across the levels of the fixed factor (McLean *et al.* 1991). The alternative formulation of EMS only changes the expected value of the mean square for the random block effect anyway, although the interpretation of the blocks term in these ANOVAs still creates considerable debate among statisticians (Samuels *et al.* 1991 and subsequent comments in same issue). Note that the EMS for the non-additive model are identical to those derived for the two factor crossed model described in Chapter 9.

10.2.3 Null hypotheses

There are two null hypotheses of interest in RCB (or simple RM) designs. The most important is the test for treatment effects, but the test of block effects might also be of some interest. The statistical tests of these null hypotheses depend on the expected mean squares (EMS) which in turn depend on whether we consider an interaction likely and whether the factors (treatments and blocks) are considered fixed or random. The most common situation in biological experiments is where block or subject is a random factor (the blocks used in the experiment are a random sample from a larger population of blocks and we wish to generalize our results to this population of blocks) and factor A ("treatment") is fixed, although other combinations are possible.

Factor A (fixed)

$H_0(A)$: $\mu_1=\mu_2=\ldots=\mu_i=\ldots=\mu_p$. This H_0 states that there is no difference between the factor A marginal means, pooling blocks. Using the experiment from Walter & O'Dowd (1992), the H_0 is no difference between the mean number of mites per leaf for the two domatia treatments, pooling leaf pairs (blocks).

This is equivalent to:

H_0: $\alpha_1=\alpha_2=\ldots=\alpha_i=0$, i.e. no effect of any level of factor A and therefore all treatment effects equal zero. For this example, there is no effect of domatia treatment on the mean number of mites per leaf (Walter & O'Dowd 1992).

Blocks (random)
H_0(B): $\sigma_\beta^2 = 0$, i.e. no added variance due to all possible blocks that could have been used. From Walter & O'Dowd (1992), the H_0 is that there is no added variance due to all possible leaf pairs that could have been used.

***F*-ratios**
We can test these null hypotheses by seeing which of our mean squares have the same expected value when the H_0 is true. The *F*-ratio from these mean squares will follow an *F* distribution if certain assumptions (see Section 10.4) hold. It is clear from Table 10.5 that MS_A and $MS_{Residual}$ have the same expected value when there is no effect of factor A so these two mean squares are used in an *F*-ratio to test the H_0. Note that when there are only two levels of factor A (treatments), the *F* test for treatments in a RCB or RM design is equivalent to a paired *t* test (Chapter 3) of the H_0 that the mean of the paired treatment differences equals zero.

There is no test for block effects if we allow for an interaction between treatments and blocks when blocks are random because there is no MS that has the same expected value as MS_{Blocks} when H_0 is true (the $MS_{Residual}$ has two components, rather than just σ_ε^2). As we pointed out in the previous section, the hypotheses being tested about block effects depend on which version of the EMS we use for mixed models (Chapter 9) and is an issue of debate among statisticians (Samuels *et al.* 1991).

For other combinations of fixed and random factors/blocks, tests of null hypotheses depend on whether we are willing to assume an underlying additive (no treatment by block interaction) model or not. For example, if both factor A (treatments) and blocks are fixed, there is no test for either unless we assume an underlying additive model. If factor A is random (H_0: $\sigma_\alpha^2 = 0$) but blocks are fixed (an unusual combination in practice), there is also no test for A unless we assume an additive model.

The complete ANOVAs and interpretation of the studies from Walter & O'Dowd (1992) and Driscoll & Roberts (1997) are presented in Box 10.1 and Box 10.2 respectively. These are both mixed models with blocks (or subjects) random, so the tests for blocks are only valid if there is no A by blocks interaction.

10.2.4 Comparing ANOVA models

The SS, df and MS for each term in the classical ANOVA (Table 10.2) can also be derived from comparing the fit of a full and a reduced linear model, where the reduced model simply omits the parameter specified to be zero in the H_0 – see Box 10.4. The approach of comparing linear models also offers strategies for handling missing values (Section 10.9) whereas the formulae in Table 10.2 are only applicable when there are no missing values. We use SS (and MS) to measure the fit of the different models, although likelihoods could also be used with the likelihood ratio replacing the *F*-ratio for testing whether the reduced model fits significantly worse than the full model (Chapter 13).

10.3 Interactions in RCB and RM models

10.3.1 Importance of treatment by block interactions

If we do assume an underlying additive model, with its associated expected mean squares, then the H_0 for factor A is that there is no effect of treatments in any block (Newman *et al.* 1997). The additive model also allows a test of the block effects, although these are usually not of much interest in practice except for determining the efficiency of blocking compared to a CR design – see Section 10.7. Also, assuming an additive model allows a test of factor A and blocks if both are fixed; neither is testable in the presence of an interaction using the non-additive model. How realistic is the additive (no interaction) model for biological experiments?

Newman *et al.* (1997) argued that if blocks are spatial units defined by the investigator, then we might consider them simply large, randomly chosen, experimental units, and hence treatment by block interactions are unlikely. In contrast, others (e.g. Mead 1988, Underwood 1997) have argued that factor A by block interactions are quite likely in biological experiments, particularly for field experiments where experience suggests that treatment effects may vary spatially. Newman *et al.* (1997) suggested that interactions are more likely when blocks are naturally occurring units (e.g. organisms, genotypes) than when the scale of blocks is chosen by the investigator.

Table 10.5 Expected mean squares (EMS) for additive and non-additive models for a RCB design. There are *p* levels of factor A (treatments) and *q* blocks

Additive model

Source	A fixed, blocks fixed	A random, blocks random	A fixed, blocks random	*F*-ratio
A	$\sigma_\varepsilon^2 + q\frac{\sum_{i=1}^{p}\alpha_i^2}{p-1}$	$\sigma_\varepsilon^2 + q\sigma_\alpha^2$	$\sigma_\varepsilon^2 + q\frac{\sum_{i=1}^{p}\alpha_i^2}{p-1}$	$\frac{MS_A}{MS_{Residual}}$
Blocks	$\sigma_\varepsilon^2 + p\frac{\sum_{j=1}^{q}\beta_j^2}{q-1}$	$\sigma_\varepsilon^2 + p\sigma_\beta^2$	$\sigma_\varepsilon^2 + p\sigma_\beta^2$	$\frac{MS_{Blocks}}{MS_{Residual}}$
Residual	σ_ε^2	σ_ε^2	σ_ε^2	

Non-additive model

	A fixed, blocks fixed		A random, blocks fixed		A random, blocks random	
Source	EMS	*F*-ratio	EMS	*F*-ratio	EMS	*F*-ratio
A	$\sigma_\varepsilon^2 + q\frac{\sum_{i=1}^{p}\alpha_i^2}{p-1}$	No test	$\sigma_\varepsilon^2 + q\sigma_\alpha^2$	No test	$\sigma_\varepsilon^2 + \sigma_{\alpha\beta}^2 + q\sigma_\alpha^2$	$\frac{MS_A}{MS_{Residual}}$
Blocks	$\sigma_\varepsilon^2 + p\frac{\sum_{j=1}^{q}\beta_j^2}{q-1}$	No test	$\sigma_\varepsilon^2 + q\sigma_\beta^2$	No test	$\sigma_\varepsilon^2 + \sigma_{\alpha\beta}^2 + p\sigma_\beta^2$	$\frac{MS_{Blocks}}{MS_{Residual}}$
Residual	$\sigma_\varepsilon^2 + \frac{\sum_{i=1}^{p}\sum_{j=1}^{q}(\alpha\beta)_{ij}^2}{(p-1)(q-1)}$		$\sigma_\varepsilon^2 + \sigma_{\alpha\beta}^2$		$\sigma_\varepsilon^2 + \sigma_{\alpha\beta}^2$	

Box 10.4 Fitting general linear models to test factor A in RCB design from Walter & O'Dowd (1992)

Full model fitted:

$(\text{Log mite number})_{ij} = \mu + (\text{treatment})_i + (\text{block})_j + \varepsilon_{ij}$

	SS	df	MS
Explained	54.399	14	3.886
Unexplained	36.007	13	2.770
$r^2 = 0.602$			

Reduced model fitted:

$(\text{Log mite number})_{ij} = \mu + (\text{block})_j + \varepsilon_{ij}$

	SS	df	MS
Explained	23.058	13	1.774
Unexplained	67.348	14	4.811
$r^2 = 0.255$			

Difference in fit of two models:

Full $SS_{Explained}$ (54.399) – Reduced $SS_{Explained}$ (23.058) = 31.341 with 1 df

$MS_A = 31.341$, which is MS_A from randomized block ANOVA (see Box 10.1).

Test of A:

$F = MS_A$ / Full $MS_{Residual} = 31.341 / 2.770$ with 1,13 df = 11.32, $P = 0.005$.

Fortunately, the test for treatments ($MS_A/MS_{Residual}$) is statistically valid for a mixed model (A fixed, blocks random), whether we assume an additive model or not. If we allow for an interaction between A and block by using the expected mean squares from the non-additive model (Table 10.5), both MS_A and $MS_{Residual}$ include $\sigma_\varepsilon^2 + \sigma_{\alpha\beta}^2$ in their expectations, the variance due to random differences between observations within each cell and the variance due to the interaction between treatments and blocks. With only n equals one per cell, we cannot separately estimate these two variances. These expected mean squares suggest that the test for factor A is really for the presence of an effect of treatments over and above the interaction between A and blocks (which still might exist, even if we cannot measure it in our unreplicated RCB or RM experiment) and true error variation. Bergerud (1996) suggested that treatment effects over and above interaction effects would occur when the treatment rankings are consistent for each block, even if the actual differences between treatments change from block to block (a treatment by block interaction). The treatment by block interaction is only statistically critical when blocks are fixed, in which case there is no test of treatments unless we assume the A by block interaction is zero (Kirk 1995, Neter *et al.* 1996).

Even if we allow for an underlying non-additive model when determining our EMS and constructing our *F*-ratios for the mixed model case, the presence of A by block interactions can result in two other difficulties when interpreting the treatment effects. First, if there is an interaction, then the $MS_{Residual}$, whose expected value contains $\sigma_\varepsilon^2 + \sigma_{\alpha\beta}^2$, will increase proportionally more than MS_A, whose expected value also includes treatment effects. The *F*-ratio for A will therefore have relatively less power in the presence of an

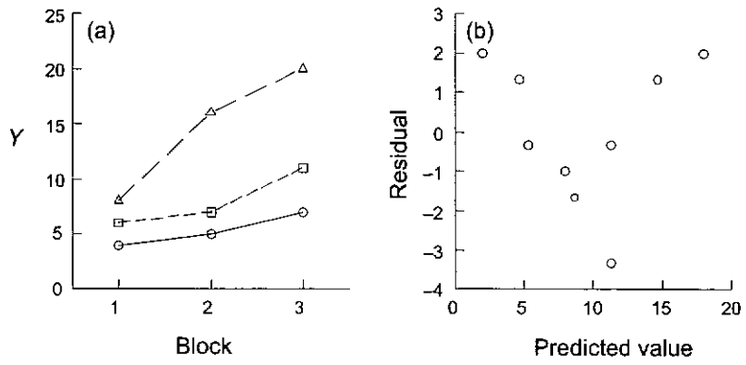

Figure 10.3 Illustration of detection of treatment by block interaction with (a) interaction plot and (b) residual plot – artificial data. Note that the difference between treatments is much greater for blocks 2 and 3 compared to block 1 but there is change in direction of treatment effects – no crossing over in interaction plot. There is clear evidence of a curvilinear relationship in the residual plot where the residuals change from positive to negative and back to positive as the predicted values increase.

interaction. Second, if the interaction is very strong, then there can be a logical difficulty with the interpretation of main effects, as discussed in Chapter 9 and by Underwood (1997). Complex interactions, where the effects of treatments are strong but in different directions between blocks, can result in a non-significant main effect of A averaging over blocks. Meaningful interpretation of such a non-significant main effect is difficult. However, interpretation of significant main effects can still be valid in the presence of an interaction (see Chapter 9). A significant main effect indicates that, averaging over blocks, there is a treatment effect even if the magnitude of that effect varies from block to block.

We recommend that you should check for A by block interactions in analyses of RCB and RM designs. Interpretation of main effects in the non-additive model may need to be constrained if strong interactions are present; the additive model, which may be necessary if blocks are fixed, relies on the absence of interactions.

10.3.2 Checks for interaction in unreplicated designs

With only one replicate experimental unit in every treatment–block combination, there is no formal test for an interaction. However, there are three ways in which an interaction between A and blocks might be detected. The first two are graphical and the third is a test for a particular type of interaction. We illustrate all three methods for the two worked examples in Box 10.1 and Box 10.2.

Cell "mean" plots

We discussed plots of cell means to interpret interactions for factorial designs in Chapter 9. These are simply plots (usually line graphs) of cell means, where the levels of one factor are used to define groups along the horizontal axis, the vertical axis is the value of the response variable and the different levels of the second factor are represented by different symbols (joined by lines). We can use the same plots for RCB or RM designs, except that the horizontal axis represents blocks or subjects and we plot the single values within each treatment–block combination (Figure 10.3(a)). Note that we still refer to population means for each cell (treatment–block combination) although we only estimate those means with n equals one in each cell in a RCB (RM) design. As with CR factorial designs, the lines should be roughly parallel if there is no interaction.

Residual plots

Another way to detect interactions is to examine the residuals. An interaction would be suggested if the pattern of the residuals changed markedly from block to block (Neter *et al.* 1996). Two graphical diagnostic techniques using residuals have been described in the statistical literature for showing interactions in RCB (RM) designs:

1. A plot of residuals against predicted values, the typical residual plot we have used extensively for assessing the adequacy of linear models in earlier chapters, is important in checking homogeneity of variance and the presence of

outliers and can also detect some types of interaction. A curvilinear relationship in this plot, where the residuals change from positive to negative and back to positive again as the predicted values increase, indicates a particular sort of block by treatment interaction (Neter *et al.* 1996), where the relative magnitudes of the treatment effects differ between blocks but not the direction of the effects (i.e. no crossing over in interaction plot) – see Figure 10.3(b). This is the sort of interaction that can often be removed by transformation (Box *et al.* 1978). In contrast, complex interactions where the direction of treatment effects changes between blocks are not easily detected with residual plots.

2. A plot of residuals against estimated comparison values (Emerson & Hoaglin 1983), where each comparison value is $(\alpha_i\beta_j)/\mu$ for each cell *ij* from the fit of the additive model 10.1. The estimates of μ, α_i and β_j are described in Box 10.3. Any consistent pattern suggests the presence of an interaction. Emerson & Hoaglin (1983) argued that this plot is particularly useful for determining the strength of an interaction already suggested by the first residual plot or a cell means plot and helps choose a transformation that might restore additivity. If the slope of the best-fit line on this plot is k, then a power transformation using a power of $1-k$ will be effective. Emerson & Hoaglin (1983) also recommended using robust estimates of effects for calculating comparison values, such as those from a median polish (see Section 10.5), to distinguish systematic non-additivity from the effects of just one or two unusual values.

Tukey's test for (non-)additivity

Tukey (1949) developed a test to detect one particular type of interaction in unreplicated factorial designs. Tukey's test for additivity can be viewed as a test of the curvilinear relationship between the residuals and the predicted values from the original linear model (Box *et al.* 1978), the relationship we were trying to detect with the residual plot described above. It is also a specific contrast–contrast test on the interaction (Hays 1994, Kirk 1995) where the contrast coefficients are $(\bar{y}_i - \bar{y})$ and $(\bar{y}_j - \bar{y})$. Kirk (1995) pointed out that Tukey's test for additivity is best at detecting relatively simple interactions which involve different magnitudes of treatment effects for each block but not different directions of the treatment effects (i.e. lines in interaction plot are not parallel but do not cross). He also suggested that a liberal significance level should be used ($\alpha = 0.10$ or 0.25) to reduce the risk of a Type II error (not detecting a real interaction), a recommendation we support.

The computational details are provided in Box 10.5 and illustrated using the data from Driscoll & Roberts (1997). Basically the $SS_{Residual}$ is split into that due to the specific type of non-additivity described above and that remaining. This $SS_{non-add}$ is a single df component from the original $SS_{Residual}$ and the remaining $(q-1)(p-1)-1$ df

Box 10.5 Tukey's test for (non-)additivity, illustrated for data from Driscoll & Roberts (1997)

Recall the non-additive linear model from Box 10.3 for the RCB/RM design:

$$y_{ij} = \mu + \alpha_i + \beta_j + [(\alpha\beta)_{ij}] + \varepsilon_{ij}$$

We can redefine $(\alpha\beta)_{ij}$ as $D\alpha\beta_i$ where D is a second-order polynomial function of α_i and β_i and represents the multiplicative relationship between factor A and blocks (Neter *et al.* 1996, Sokal & Rohlf 1995). The value of D is, using the terminology of Neter *et al.* (1996):

$$D = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q}\alpha_i\beta_j y_{ij}}{\sum_{i=1}^{p}\alpha_i^2\sum_{j=1}^{q}\beta_j^2}$$

where α_i and β_j are the effects of factor A and blocks respectively, as defined in Box 10.3. We replace these parameters by their sample estimates to obtain the estimated value of D:

$$\hat{D} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{y}_i - \bar{y})(\bar{y}_j - \bar{y})y_{ij}}{\sum_{i=1}^{p}(\bar{y}_i - \bar{y})^2\sum_{j=1}^{q}(\bar{y}_j - \bar{y})^2}$$

The SS for this specific form of non-additivity is $\sum_{i=1}^{p}\sum_{j=1}^{q} D^2\alpha_i^2\beta_j^2$ and this is estimated by:

$$\sum_{i=1}^{p}\sum_{j=1}^{q}\hat{D}^2(\bar{y}_i - \bar{y})^2(\bar{y}_j - \bar{y})^2$$

which equals:

$$\frac{\left[\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{y}_i - \bar{y})(\bar{y}_j - \bar{y})y_{ij}\right]^2}{\sum_{i=1}^{j}(\bar{y}_i - \bar{y})^2\sum_{j=1}^{q}(\bar{y}_j - \bar{y})^2}$$

To illustrate from Driscoll & Roberts (1997), here are the raw data and marginal means.

Block	1992	1993	1994	Block means
logging	4	17	18	13.00
angove	−10	−1	8	−1.00
newpipe	−15	−10	1	−8.00
oldquinE	−14	−11	−2	−9.00
newquinW	−4	6	0	0.67
newquinE	0	5	1	2.00
Year means	−6.50	1.00	4.33	−0.389

Using the equation above:

$$[\Sigma\Sigma(\bar{y}_i - \bar{y})(\bar{y}_j - \bar{y})y_{ij}]^2 = [(13-(-0.389))(-6.50-(-0.389))(4) + (-1-(-0.389))(-6.5-(-0.389))(-10) + \cdots + (2-(-0.389))(4.33-(-0.389))(1)]^2 = 510.34$$

$$\Sigma(\bar{y}_i - \bar{y})^2\Sigma(\bar{y}_j - \bar{y})^2 = (61.54)(318.53) = 19\,602.34$$

$SS_{non-add} = 510.34/19602.34 = 0.026$ with 1 df

$MS_{non-add} = 0.026$

$SS_{Remainder} = SS_{Total} - SS_A - SS_B - SS_{non-add} =$
$1516.278 - 369.444 - 955.611 - 0.026 = 191.197$ with 9 df,

$MS_{Remainder} = 191.197/9 = 21.244$

$F_{non-add} = 0.026 / 21.244 = 0.001$ with 1 and 9 df, $P = 0.974$.

No evidence of strong interaction between blocks and years, even using a liberal α of 0.25.

component represents other sorts of interaction and the variation between experimental units ($SS_{Remainder}$). These SS are converted to MS and an *F*-ratio constructed which is $MS_{non-add}$ / $MS_{Remainder}$; this *F*-ratio follows an *F* distribution and the H_0 of no interaction can be tested in the usual manner.

Additivity and transformations

If evidence of an interaction is detected, there is an argument that we should try and reduce the effect of such an interaction, as this will increase the power and interpretability of the test for treatments. Presumably, a factor A by block interaction is not important to us biologically or else we would have replicated each treatment–block combination as a generalized RCB design (see Section 10.12). If the non-additivity is due to the scale on which the response variable is measured and therefore a multiplicative relationship between the response variable and treatments and blocks, then a transformation to a different scale of measurement (e.g. logs) may remove the interaction and make the relationship additive (Chapter 9). This is the type of non-additivity Tukey's test and residual plots are likely to detect, so a significant result from Tukey's test would suggest a transformation will reduce the extent of the interaction.

10.4 Assumptions

10.4.1 Normality, independence of errors

We have already discussed the "assumption" of no factor A by block interaction, pointing out that the presence of an interaction does not invalidate the test for treatments if block is a random factor. In addition, the usual assumption that experimental units are randomly sampled from a population of experimental units is still important. We also assume, as usual, that the residuals are normally distributed and have constant variance within treatments across blocks (homogeneity of variance assumption). Plots of residuals, both within treatments and against predicted values, are interpreted in the same way as described in Chapters 8 and 9; watch out for wedge-shaped patterns suggesting an underlying skewed distribution. If the RCB or RM design is a mixed model with random blocks, the common scenario in biology, then the homogeneity of variance assumption can be incorporated into a more general assumption about variances and covariances (Section 10.4.2). Outliers from the fitted model are as important to detect for RCB (RM) designs as for CR designs. Observations with large residuals can be identified from residual plots and most statistical software will warn of outliers when the model is fitted.

Even in RCB designs and RM designs, we assume that the residuals are independent of each other, even though the observations within a block or subject are not (Kirk 1995). This is because we assume that block effects are independent of residual effects, an assumption which is justified by the random allocation of levels of factor A to experimental units within a block (Brownie *et al.* 1993) or the random order of treatment application within a subject. Spatial heterogeneity between experimental units within blocks can be modeled as part of the analysis (Brownie *et al.* 1993), which may increase the precision of treatment means and the power of tests of treatment effects. Note that even though we acknowledge that observations from experimental units within a block are possibly correlated, sensible interpretation of biological experiments usually relies on the experimental units within blocks being far enough apart so that the effect of one treatment doesn't affect any other experimental unit, e.g. animals crawling off one experimental unit in response to a treatment and onto another. Similarly, in repeated measures designs, carryover effects must be explicitly avoided (by randomizing order of treatments and/or leaving a long enough gap between treatments) or be explicitly incorporated into the design and the hypotheses (see Kirk 1995).

10.4.2 Variances and covariances – sphericity

We have already indicated that in two factor linear models where one factor is random, the observations from the same level of the random factor are correlated with each other (Chapter 9). This correlation is exacerbated in RCB designs, because the experimental units in a block are often located close together, and in RM designs, because we have repeated observations on the same subject. This implies that the observations within a block, i.e. the observations from different treatments within a block (or within a subject in the repeated measures context) are not independent of each other (Kirk 1995). Therefore, we not only have to be concerned about variances in these analyses but also about covariances (correlations). These variances and covariances can be expressed in the form of a variance–covariance matrix (see Chapter 15) whose diagonal matrix contains the variances between observations within each treatment and the other entries are the covariances between treatments (i.e. the covariances between observations from different treatments).

There are two conditions that must be met for the *F*-ratio for factor A to follow an *F* distribution when we fit a two factor mixed ANOVA model to data from a RCB or simple RM design. Not only do the variances have to be the same across treatments (the usual homogeneity of variance assumption) but the covariances (i.e. the correlations between treatments within each block or subject) also have to be the same. If the variances are all equal and the covariances are all equal, i.e. the correlations between all pairs of treatments are equal, then the variance–covariance matrix shows compound symmetry. This is a *sufficient* condition for the *F*-ratio to follow an *F* distribution but it is too restrictive an assumption, i.e. it is not a *necessary* condition. The *F*-ratio for factor A in the analyses of mixed model RCB and RM designs will follow an F distribution if the variance–covariance matrix shows a pattern known as sphericity. Put simply, the sphericity condition is that the variances of the differences between values of the response variable are the same for all pairs of treatments (see Box 10.6). The sphericity assumption is

Box 10.6 Illustration of compound symmetry and sphericity assumptions using data from Driscoll & Roberts (1997)

Compound symmetry assumption:

$\sigma^2_{11} = \sigma^2_{22} = \sigma^2_{33}$ and $\sigma_{21} = \sigma_{31} = \sigma_{32}$, i.e. treatments variances are equal and treatment covariances are equal.

	General covariance matrix			Specific covariance matrix		
	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3
Year 1	σ^2_{11}			59.90		
Year 2	σ^2_{21}	σ^2_{22}		79.40	113.20	
Year 3	σ_{31}	σ_{32}	σ^2_{33}	34.80	57.80	56.27

Estimates from data suggest difference between variances and between covariances.

Sphericity assumption:

$\sigma^2_{1-2} = \sigma^2_{1-3} = \sigma^2_{2-3}$, i.e. variances of differences between treatments are equal.

	Year 1	Year 2	Year 3	Year 1–2	Year 1–3	Year 2–3
Block 1	4	17	18	−13	−14	−1
Block 2	−10	−1	8	−9	−18	−9
Block 3	−15	−10	1	−5	−16	−11
Block 4	−14	−11	−2	−3	−12	−9
Block 5	−4	6	0	−10	−4	6
Block 6	0	5	1	−5	−1	4
s^2				14.30	46.57	53.87

The estimates of the variances of the treatment differences vary, with the variance of the year 1 minus year 2 difference considerably smaller than the other two differences, a strong indication that sphericity is not met.

much less restrictive than compound symmetry because it does not require equality of variances and equality of covariances. Note that compound symmetry is simply one form of sphericity; a variance–covariance matrix which shows compound symmetry also shows sphericity by definition. If the sphericity assumption is not met, then the F test for treatments in RCB and RM designs can be liberal, i.e. the actual Type I error rate can exceed the nominal rate we set with our *a priori* significance level (Boik 1979, Box 1954). The F test is not very robust to this assumption.

There is no reason to expect the variances of the differences between pairs of treatments to be very different in classical RCB designs because the treatments are randomly allocated to different experimental units within each block. Think of this in terms of treatment correlations – the correlation between treatments one and two should not be very different from the correlation between treatments two and three if the experimental units are randomly arranged in each block and each experimental unit is randomly allocated to a treatment. In contrast, the sphericity assumption is less likely to hold for RM designs because observations for repeated measurements closer together in time will probably be more correlated than for repeated measurements further apart in time. If the order in which the treatments are applied to each experimental unit (subject) is randomized (treatments × subjects designs), then correlations between treatments might still be similar. However, in subjects by trials designs where the treatments are times (or time intervals), we would expect quite different correlations between times closer together compared to those further apart.

Note that the assumption of compound symmetry, or the more realistic assumption of sphericity, of the variance–covariance matrix only applies to mixed model RCB and RM analyses. If both factor A and blocks (or subjects) are fixed, then the linear model implies that observations are uncorrelated within treatments and within blocks (or subjects). This is probably why few textbooks (but see Kirk 1995, Neter *et al.* 1996) discuss any requirement for specific patterns of variances and covariances for RCB designs – such designs are usually presented with fixed blocks (e.g. Hocking 1996). In contrast, RM designs are nearly always presented with subjects as random and hence the pattern of variances and covariances receives considerable attention. Note also that if there are only two treatments, then sphericity is not relevant because the variance–covariance matrix is actually a vector (only two variances and a single covariance).

There are two broad approaches for dealing with violations of the assumption of sphericity, adjusting univariate F tests to make them more conservative or using a multivariate test that does not assume sphericity.

Adjusting univariate F tests

The degree to which the variance–covariance matrix departs from compound symmetry and sphericity is measured by the epsilon (ε) parameter (Winer *et al.* 1991, Keselman & Keselman 1993, Kirk 1995). When sphericity is met, ε equals one; the further ε is from one, the more the sphericity assumption is violated. An estimate of ε can be determined from the sample variance–covariance matrix and is termed the Greenhouse–Geisser epsilon ($\hat{\varepsilon}$); it is complex to calculate, requiring some matrix gymnastics. The df for the F test for factor A can then be adjusted downwards based on the value of $\hat{\varepsilon}$:

$$\mathrm{df}_{\mathrm{adj}} = (p-1)\hat{\varepsilon} \text{ and } (p-1)(q-1)\hat{\varepsilon} \qquad (10.6)$$

and the F test based on these adjusted df approximately follows an F distribution even when sphericity is not met. Unfortunately, the Greenhouse–Geisser estimate of ε can be conservatively biased when ε is close to 0.75 (Collier *et al.* 1967; see also Keselman & Keselman 1993, Winer *et al.* 1991), i.e. the adjustment to the df is too severe, making the test too conservative. An alternative estimate of ε is the Huynh–Feldt epsilon, although this can exceed one and therefore might be too liberal. Both estimates of ε and adjustments to df are standard output from most statistical software and we recommend the Greenhouse–Geisser adjustment because the true value of ε is never known so it is difficult to decide when to use the Huynh–Feldt version.

Note that a simpler version of the Greenhouse–Geisser adjustment is to set $\hat{\varepsilon}$ to its smallest value, which depends on the number of treatment groups (p) and equals $1/(p-1)$ – see Kirk (1995). This saves having to calculate $\hat{\varepsilon}$ but will obviously be conservative (it sets $\hat{\varepsilon}$ to the minimum value irrespective of the actual value of ε) and is unnecessary since most statistical software will calculate $\hat{\varepsilon}$.

In the Driscoll & Roberts (1997) example, the Greenhouse–Geisser epsilon was 0.71 and the Huynh–Feldt epsilon was 0.92, supporting the argument that the former is a more conservative estimate of ε. The adjusted df produced more conservative P values but did not change our conclusions about the effect of years (Box 10.2). There were only two treatments in the Walter & O'Dowd (1992) study, and therefore only one covariance, so the sphericity assumption was not relevant.

Multivariate tests

Another approach to dealing with the sphericity assumption is to use a procedure that does not require this assumption. We could use the differences between pairs of treatments (e.g. between pairs of times) as multiple response variables in a multivariate ANOVA (MANOVA; see Chapter 16). If there are p treatments (e.g. times), then only $p-1$

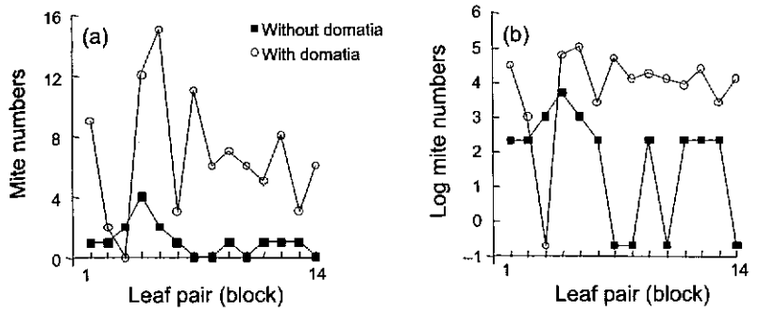

Figure 10.4 "Interaction" plots, with mite numbers plotted against block for the two treatments, for untransformed (a) and transformed (b) Walter & O'Dowd (1992) data.

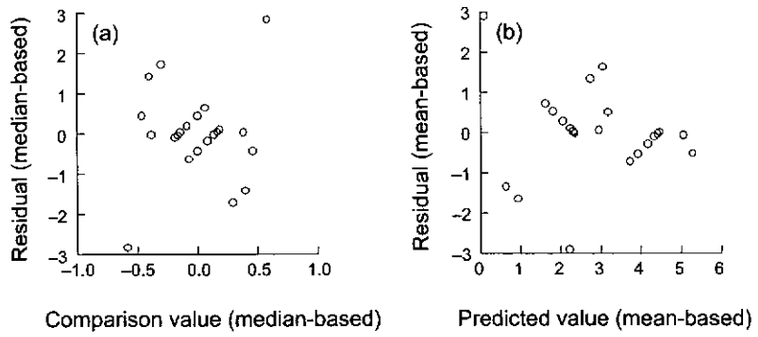

Figure 10.5 (a) Plot of residuals against comparison values from a median polish from Walter & O'Dowd (1992) data. (b) Plot of residuals against predicted values (mean-based) from Walter & O'Dowd (1992) data.

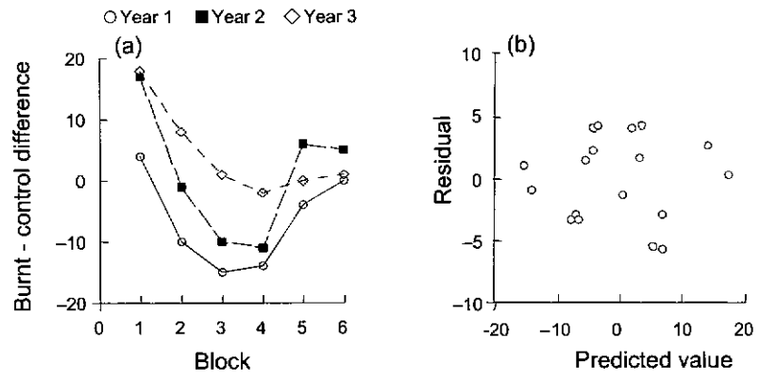

Figure 10.6 (a) "Interaction" plot, with burnt–control differences plotted against block (catchment) for each year. (b) Plot of residuals versus predicted values, for the Driscoll & Roberts (1997) data.

differences need to be used. The H_0 is that the population mean of the differences for all pairs of treatments equals zero. Because we are testing two or more population means simultaneously (e.g. with p equals three, there are two differences), we are really testing whether the $p-1$ differences have a population mean vector equal to zero (Keselman & Keselman 1993). Any of the test statistics used in MANOVA are applicable here but as discussed in Chapter 16, we recommend the Pillai trace statistic.

The MANOVA approach does not assume sphericity of the variance–covariance matrix but does assume multivariate normality, which is always difficult to check. It also requires more subjects or blocks than treatments, otherwise the MANOVA will encounter computational difficulties. In the Driscoll & Roberts (1997) example, the P value from the MANOVA testing whether the population mean differences between all pairings of the three years equals zero was 0.016, leading us to the same conclusion as for the adjusted univariate analysis (Box 10.2).

10.4.3 Recommended strategy

Formal tests of sphericity include Mauchley's test, which is very sensitive to deviations from multivariate normality and is not recommended (Keselman & Keselman 1993), and the "locally best invariant test", which is tedious to calculate (Kirk 1995). We suggest, like others (Keselman & Keselman 1993, Winer *et al.* 1991), that it is probably safer to assume that this assumption is not met and use adjusted univariate F-ratios or the multivariate approach (see Looney & Stanley 1989, Manly 1992, Potvin *et al.* 1990, von Ende 1993). Which is the best approach? As usual in applied statistics, that depends on the nature of the data. Looney & Stanley (1989) suggested using both approaches (most statistical packages automatically provide both analyses); if either the adjusted univariate or the multivariate indicates a significant result, reject the H_0. If neither indicate a significant result, do not reject H_0. Most statistical software routinely outputs all three approaches (unadjusted univariate, adjusted univariate, multivariate).

10.5 Robust RCB and RM analyses

The only commonly used robust alternatives to the analyses we have described in this chapter are to transform the observations to ranks and then do the usual parametric analysis on the ranked data. The ranking can be done in two ways.

- Rank the data separately within each block (or subject) and then use the usual F test for factor A described earlier in this chapter. Note that this test is equivalent to the Friedman test (Hollander & Wolfe 1999), which also ranks the observations within each group but compares its test statistic to a chi-square distribution. The Friedman test is an extension of the Kruskal–Wallis test described in Chapter 8 for single factor ANOVA models.
- Alternatively, the data could be ranked over the entire data set as described for rank-transform (RT) procedures in CR designs (Chapters 8 and 9) and the usual F test for treatments applied to the ranked data (see Maxwell & Delaney 1990).

All of our previous comments about rank-based analyses (see Chapters 3 and 8) apply here, particularly that these tests do not assume normality but do not necessarily solve problems about variances and covariances (Maxwell & Delaney 1990) and can be inefficient when there are many ties (Neter *et al.* 1996). Rank-based tests do not deal with interactions very well (Chapter 9) so it is difficult to predict what effect block by treatment interactions will have on the analysis.

An alternative approach is to estimate the effects of the linear model in a robust manner, i.e. obtain estimates of the factor A and block (or subject) effects that are not sensitive (i.e. are resistant) to outliers. Emerson & Hoaglin (1983) and Emerson & Wong (1985) proposed a technique called a median polish, which uses the medians to fit an additive model of the form:

$$y_{ij} = m + \alpha_i + \beta_j + \varepsilon_{ij} \qquad (10.7)$$

where m is the overall median and α_i, β_j and ε_{ij} are factor A effects, block effects and residuals estimated using marginal medians instead of marginal means. Median polish determines these effects in an iterative fashion, calculating the row effects, then the column effects, then recalculating the row effects, etc. The computations are a little tedious (see Emerson & Hoaglin 1983; MINITAB™ also provides median polish) but the results are useful for detecting some forms of non-additivity, by using comparison values calculated from the median polish (Section 10.3.2), and also for providing more robust detection of outliers.

Randomization tests are also possible by generating the distribution of an appropriate test statistic by randomly reallocating observations to treatment–block combinations, as we described for factorial designs in Chapter 9 (Manly 1997).

Finally, if the response variable being analyzed has a known distribution that fits an exponential form, then generalized linear modeling procedures can be used (Chapter 13). GLMs measure the fit of models with maximum likelihood techniques, allow a variety of underlying distributions, such as Poisson, binomial, lognormal, etc., and tests of hypotheses about model parameters use likelihood ratios.

10.6 Specific comparisons

Planned contrasts and unplanned multiple comparisons between factor A levels in RCB (or RM) designs depend on whether the sphericity assumption is met, because these tests usually rely on a single error term, the $MS_{Residual}$. We argued in Section 10.4.2 that for classical RCB designs, where treatments are randomly allocated to independent experimental units within blocks, the sphericity assumption is less likely to be violated. The usual contrasts and pairwise comparison procedures described in Chapters 8 and 9 can be used; $MS_{Residual}$ would be used as the error term for calculating the standard errors of these comparisons. For RM designs, the variances and covariances are less likely to conform to sphericity (and adjusted df cannot easily be calculated for specific comparisons) so we agree with Kirk (1995) that separate denominators should be used for each pairwise (or more complex) comparison. Keselman & Keselman (1993) proposed pairwise t tests with separate error terms based on the two levels being compared. For example, to compare groups 1 and 2 for factor A:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2 + s_2^2 - 2s_{12}}{q}}} \qquad (10.8)$$

where s_1^2 and s_2^2 are the sample variances for groups 1 and 2, s_{12} is the sample covariance between groups 1 and 2 and q is the number of subjects or blocks. Note that this is simply a paired t test (see Chapter 3) comparing the means of the two groups. The SS_A can also be partitioned into SS for each comparison, as described in Chapter 8 and the two groups compared with an F test. Not all statistical software provides separate denominators for these F-ratio tests of each contrast, so we illustrate the calculations of separate error terms for the Driscoll & Roberts (1997) data in Box 10.7. The calculated F-ratio statistic will be the same as the t statistic. For unplanned pairwise comparisons where control over the familywise Type I error rate is required, a Bonferroni-type correction (see Chapter 3) can be applied.

Trends (linear, quadratic, etc.) across levels of factor A can also be tested using the methods outlined in Chapter 8; these tests are often default output from some statistical software if the data are coded as repeated measures. The only difference in testing for trends between a RCB (or RM) design and a CR design is which denominator to use for the F-ratio (Kirk 1995). As for other contrasts described above, separate error terms for each trend test should be used if sphericity might not hold. Note that Winer *et al.* (1991) also suggested using a separate denominator for each trend (linear, quadratic, etc.), although their tests for each trend component are only slightly more conservative than those based on the $MS_{Residual}$ and we prefer the approach of Kirk (1995).

10.7 Efficiency of blocking (to block or not to block?)

The decision to include a blocking factor in an experimental design depends on two questions.

- Are experimental units in blocks more similar to each other than to other experimental units

Box 10.7 Calculation of separate error term for contrasts in RCB/RM analyses (see Kirk 1995), using the data of Driscoll & Roberts (1997)

First, calculate the relevant contrast (trend or otherwise) for each block/subject, e.g. the linear trend for block 1:

$\psi = (-1)4 + (0)17 + (1)18 = 14$, where -1, 0 and 1 are the contrast coefficients (c_i) for a linear trend through three equally spaced levels (see Table 8.8).

Second:

(i) sum the trend values across blocks/subjects ($\Sigma\psi$) = 65
(ii) sum the squared trend values across blocks/subjects ($\Sigma\psi^2$) = 937
(iii) sum the squared contrast coefficients (Σc_i^2) = 2

$$SS_{Residual(linear)} = \frac{\sum_{j=1}^{q}\psi^2 - \frac{\left(\sum_{j=1}^{q}\psi\right)^2}{q}}{\sum_{i=1}^{p}c_i^2} = [(937) - (65)^2/6]/2 = 116.42 \text{ with } q-1 = 5 \text{ df.}$$

$$MS_{Residual(linear)} = SS_{Residual(linear)}/q - 1 = 116.42/5 = 23.28$$

in different blocks? If so, then the blocking factor will explain some of the residual variation, resulting in a smaller $MS_{Residual}$ and a more powerful test of factor A.
- Does the reduction in $MS_{Residual}$ compensate for the loss of df in the ANOVA?

In most cases, the answer to the second question is unknown unless good pilot data are available, so a decision to block is made primarily on the likely extent of between-block variation.

After an RCB experiment, we might also wish to know whether using blocks was a better experimental design than a CR experiment without any blocking, in terms of precision of estimates of treatment effect and power of the tests for factor A. Lentner *et al.* (1989) argued that a measure of relative efficiency (RE) should be used to compare an RCB design to a CR design, where RE is defined as the ratio of the variance of the treatment comparison of the CR experiment to the variance of the treatment comparison in the RCB experiment. Larger REs indicate that the RCB design produced a more precise (lower variance) estimate of treatment effects compared with the CR design. Based on the work of Yates and Kempthorne, they defined an estimate of this relative efficiency (ERE) for a RCB design compared with a CR design:

$$ERE = \frac{(q-1)MS_{Block} + q(p-1)MS_{Residual}}{(pq-1)MS_{Residual}} \quad (10.9)$$

Lentner *et al.* (1989) also noted that ERE is monotonically related to the ratio of $MS_{Blocks}/MS_{Residual}$, even though that *F*-ratio is inappropriate for testing blocks in the non-additive model, so either ERE or the *F*-ratio could be used. If either is greater than one then an RCB design is more efficient than a CR design where the number of replicates per treatment is equal to the number of blocks. In the Walter & O'Dowd (1992) example, the *F*-ratio for blocks is less than one so the RCB design probably did not offer more efficiency than a CR design in this case. The *F*-ratio for subjects (blocks) was much greater than one in the Driscoll & Roberts (1997) example, although it is difficult to envisage how such an RM design could have been set up as a CR design, so the efficiency of blocking is not so relevant.

10.8 Time as a blocking factor

In all the examples we have so far used in this chapter, the blocks have been spatial units, locations in space. There are occasions where you only have a small number of experimental units and can only afford to have a single replicate of each treatment in an experiment (experimental units might be very large or very expensive). One option in these situations might be to repeat the unreplicated experiment a number of times. The experiment can then be analyzed as an RCB design with time as a blocking factor. One problem with using time as blocks is deciding whether time is a random factor. If we run the experiment over three successive weeks, for example, it's difficult to imagine from what population of times these three are a random sample. Under these circumstances, time might be treated as a fixed factor, which restricts us to using the additive model and therefore assuming no factor A by block interactions. Alternatively, we could argue that we have a random sample of at least a month or two, so time is random, and we are no worse off in our generalization – doing the experiment as a completely randomized one factor design would take only one week, and we couldn't generalize that result to any other time, anyway.

One particular design that uses time as a blocking factor is a crossover design, described in Section 10.11.4.

10.9 Analysis of unbalanced RCB designs

Missing observations are potentially a big problem for RCB and RM designs because a single missing observation is, in effect, a missing cell. The equations in Table 10.2 are not appropriate when there are missing observations. The simplest approach to missing observations in RCB (RM) designs is to omit the whole block or subject that has the missing value(s). This is the default approach for most statistical software if data are arranged, and the analysis done, as a classical "repeated measures" ANOVA (Section 10.13). Of course, this removes non-missing observations from the block/subject with the missing observation, which is wasteful of data and reduces the power of the test for factor A.

It turns out that we can analyze unreplicated two factor designs with missing observations as long as (i) there are not too many missing values and (ii) there are no treatment by block interactions. Note that we are assuming that the observations are missing randomly. Cells may also be missing by design, because the number of available experimental units is less than the number of treatment–block combinations and hence an incomplete block design should be considered (see Section 10.11.2). There are two broad analytical approaches (Box 10.8 and Chapter 4): substitute a replacement observation or compare the fit of full and reduced linear models.

If we assume additivity, then we can predict a value for any cell using Equation 10.4. Snedecor & Cochran (1989) and Sokal & Rohlf (1995) proposed a more complex method for estimating a missing value based on treatment and block totals. Both methods use the available information from the same treatment and block in estimating the missing value and produce very similar estimated values. One df should be subtracted from the residual for each substituted value. Snedecor & Cochran (1989) indicated that the SS_A (and SS_{Blocks}; see Sokal & Rohlf 1995) is slightly biased upwards and recommended a correction, although it does not make much difference in practice. Note that any procedure for estimating a missing value in an unreplicated factorial design must assume that there are no treatment by block interactions.

Alternatively, we can use the comparison of linear models approach where SS_A and SS_{Blocks} are determined by comparing the fit of a full model versus the relevant reduced model (Section 10.2.4; Box 10.4). This is the default approach for most statistical software when the data are arranged, and the analysis done, as a classical RCB design and is termed the "regression" approach by Neter *et al.* 1996. Note that the SS are no longer orthogonal, i.e. the SS for A, blocks and residual do not add to the total SS. Generally, substituting a new value as described above and comparing full and reduced additive models will result in very similar tests for the effects of treatments (Box 10.8).

Box 10.8 | Analyzing RCB designs with a missing observation

Based on Driscoll & Roberts (1997) with one observation (newpipe in 1993) missing.

Raw data and marginal means:

Block	1992	1993	1994	Block means
logging	4	17	18	13.00
angove	−10	−1	8	−1.00
newpipe	−15		1	−7.00
oldquinE	−14	−11	−2	−9.00
newquinW	−4	6	0	0.67
newquinE	0	5	1	2.00
Year means	−6.50	3.20	4.33	0.18

To estimate the missing observation, we use Equation 10.4 to estimate the predicted value for any cell, in this case the cell with the missing observation:

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

$$3.20 + (-7.00) - 0.18 = -3.98$$

This is very similar to the new value (−3.90) from the method of Snedecor & Cochran (1989) and Sokal & Rohlf (1995). We can then substitute this value for the missing observation and fit the ANOVA model as usual for this design, subtracting one df from the residual. Note that the actual value was −10, suggesting that simply predicting the missing observation assuming additivity is not ideal in this case, even though there was not strong evidence for an interaction between blocks and years.

The results of the three different approaches for dealing with this missing observation are presented below. Note the $MS_{Residual}$ is the same when we substitute a new value and when we compare full and reduced models and the tests of factor A (year) are very similar (see Neter *et al.* 1996).

	Omit block 3				Substitute new value				Model comparison approach			
Source	df	MS	*F*	*P*	df	MS	*F*	*P*	df	MS	*F*	*P*
Year	2	136.067	7.044	0.017	2	195.097	10.295	0.005	2	192.058	10.135	0.005
Block (drainage)	4	186.767			5	174.808			5	166.217		
Residual	8	19.317			9	18.950			9	18.950		

Since we are fitting a full model with no interaction term, the results from this linear models analysis will be similar to fitting the cell means model (Box 10.3) and using specific contrasts to test the subset of hypotheses for A and blocks using only those cells with data. The model we use is a restricted means model because it assumes that all treatment by block interactions are zero. Kirk (1995) illustrates using cell means models to analyze RCB designs with missing values.

This approach of comparing full and reduced effects models can only be used because the full model is an additive one with no interaction terms. In replicated factorial designs with missing cells, interactions are presumably potentially important and we cannot use a comparison of full and reduced models that include interaction terms in these circumstances (Chapter 9). Instead, the cell means approach and a subset of testable hypotheses about interactions and main effects must be used.

What is the best way of dealing with missing values in RCB or simple RM designs? The conservative approach is omitting the incomplete block or subject; it is simple, doesn't assume additivity, and is probably reasonable if the number of remaining blocks/subjects is not too small. However, the strength of inference about the effects of treatments across blocks will be reduced because we are using fewer blocks. In many cases, each block/subject may represent such an effort so that you are unwilling to discard the data from other treatments in the problem block/subject; alternatively, the number of blocks/subjects may be small and omitting one block could reduce the size of the experiment by an appreciable amount. In this case, there is no simple recommendation for which of the two alternatives (substitution, effects model comparisons) is best, although they will usually produce similar results. Both approaches assume no treatment by block interaction. Therefore, if you must analyze a design with missing observations, it is particularly important that checks for factor A by block interactions using the available data are done (Section 10.3.2). There is a downside to the model comparison approach, especially if the experiment really is a RM design where meeting the assumption of sphericity is likely to be a problem. Most software will not provide adjusted univariate or multivariate tests when general linear models are fitted (Section 10.13). This is actually a serious problem because, like other ANOVAs, the unbalanced RCB or RM ANOVAs are more sensitive to assumptions (especially sphericity) than a balanced design (Berk 1987). We can only suggest checking sphericity after omitting the block or subject with the missing value (using repeated measures coding in your statistical software) before fitting the additive linear model to the whole data set.

We illustrate the analysis of RCB or simple RM designs with a missing observation by analyzing the Driscoll & Roberts (1997) data with the observation from the second year and the third block missing (Box 10.8). In this example, there are few blocks (only six) and omitting an entire block changes the ANOVA markedly compared to the substituting a new value determined from the available data for block 3 and year 2 or simply comparing the fit of appropriate full and reduced additive linear models to the unbalanced data.

10.10 | Power of RCB or simple RM designs

The power of RCB or simple RM designs is determined similarly to a CR single factor design (see Chapter 8) except that the sample size is the number of blocks or subjects and the residual variation will probably be smaller than for a CR design. The non-centrality parameter is defined as:

$$\lambda = \sqrt{\frac{\sum_{i=1}^{p} \alpha_i^2}{\sigma_\varepsilon^2 / q}} \qquad (10.10)$$

which can also be expressed as:

$$\phi = \sqrt{\frac{\lambda}{p}} \qquad (10.11)$$

Whether Equation 10.10 or Equation 10.11 is used depends on whether we are using power tables or curves (see Neter *et al.* 1996) or power analysis software. Ideally, we would use a pilot study to provide an estimate of σ_ε^2 (the residual variance) and then determine the number of blocks (q) required to detect a treatment effect of a given size, i.e. use power analysis for determining sample size required in the design phase of the experiment, although *post hoc* calculations of power can be carried out in the same manner as described for a CR design. Note that using power calculations to determine the number of blocks or subjects required in a RCB or RM experiment probably only makes sense when blocks are

considered random; if blocks are fixed, then their number is also fixed.

10.11 | More complex block designs

10.11.1 Factorial randomized block designs

Block designs can also be extended to include factorial experiments, where all combinations of two or more factors are included in each block (Kirk 1995). For example, Brunkow & Collins (1996) did a field enclosure experiment that examined the effects of two factors (density and variance in initial size) on various response variables (growth, dry mass, stage of metamorphosis) for larval salamanders. This was a factorial design arranged in three spatial blocks with one replicate of each combination of density and initial variation in size in each block. A second example is from Wagner & Wise (1996), who set up a factorial experiment examining the effects of density (three levels: zero, low and high) and predator reduction (two levels: control and predator reduction) on growth rates of wolf spiderlings. One replicate of each combination of density and predator reduction was located in each of four spatial blocks.

The non-additive linear model, which includes block by factor interaction terms, for the factorial RCB design with two factors (A and C) replicated at a number of blocks (B) is:

$$y_{ijk} = \mu + \alpha_i + \gamma_k + \alpha\gamma_{ik} + \beta_j + \alpha\beta_{ij} + \gamma\beta_{kj} + \alpha\gamma\beta_{ikj} + \varepsilon_{ijk} \quad (10.12)$$

where α_i is the effect of factor A, γ_j is the effect of factor C, $\alpha\gamma_{ij}$ is the interaction between factors A and C, β_k is the effect of blocks, $\alpha\beta_{ik}$, $\gamma\beta_{jk}$, and $\alpha\gamma\beta_{ijk}$ are the interactions between A, C, AC and blocks and ε_{ijk} is the residual term independent of blocks. This is Model 1 of Newman *et al.* (1997). As with all unreplicated RCB designs, we cannot estimate the residual separately from at least one interaction term, in this case the $\alpha\gamma\beta_{ijk}$ interaction.

The ANOVA table based on this non-additive model with expected mean squares is shown in Table 10.6. If blocks (B) are considered random and the other factors (A and C) are fixed (the common situation with biological experiments), then each term of interest in the model (A, C, AC) is tested against its interaction with blocks and there are no tests for blocks or its interactions. If blocks and either A or C are random, then some terms will have no appropriate *F* test, e.g. if blocks and C are random, there will be no other MS in the model with same expected value as MS_A if the H_0 of no effect of A is true. If blocks and both A and C are random, then there are no tests for either A or C. In these circumstances, we must rely on quasi F-ratios as outlined in Chapter 9 or else assume an additive model. If blocks are fixed, then there are no tests for the other factors in the non-additive model so we must assume no interactions with blocks and fit an additive model as described below.

The linear model 10.12 is the equivalent of a two factor repeated measures design where both factors are "within subjects" (Keppel 1991). We argue that terms such as "factorial randomized block" and "factorial within subjects repeated measures", while useful for describing the physical structure of the experiment, actually obscure the fundamental underlying linear model, which in this case is simply an unreplicated three factor, crossed, ANOVA model (Chapter 9). The EMS provided by Kirk (1995) for a factorial RCB with blocks random are identical to those provided by Winer *et al.* (1991) for a three factor ANOVA with one factor (blocks) random.

An alternative approach is to fit an additive model:

$$y_{ijk} = \mu + \alpha_i + \gamma_k + \alpha\gamma_{ik} + \beta_j + \varepsilon_{ijk} \quad (10.13)$$

which is Model 2 of Newman *et al.* (1997) and the one which users of factorial RCB designs often fit (e.g. Brunkow & Collins 1996, Wagner & Wise 1996). This model combines the block by A, C and A × C interactions (i.e., the three residual terms in Table 10.6) into a single residual term (Table 10.7). Although the use of this pooled error term increases the degrees of freedom in the denominators used to construct the *F*-ratios, and therefore increases the power of individual tests of A, C and A × C, there are costs. First, as the additive model implies, we have to assume that there are no interactions with blocks; this assumption is very difficult to test and, for biological experiments, might not be true in some situations

Table 10.6 | ANOVA table for a factorial randomized complete block design with factors A (p levels) and C (r levels) being fixed and B (q blocks) random

Source	Wagner & Wise (1996)	df	Expected mean square	Test (A, C fixed, blocks random)
B = Block	Block	$q-1$	$\sigma_\varepsilon^2 + D_p D_r \sigma_{\alpha\gamma\beta}^2 + D_p r\sigma_{\alpha\beta}^2 + pr\sigma_\beta^2$	No test
A	Density	$p-1$	$\sigma_\varepsilon^2 + D_q D_r \sigma_{\alpha\gamma\beta}^2 + D_q r\sigma_{\alpha\beta}^2 + pr\sigma_\alpha^2$	$\frac{MS_A}{MS_{AB}}$
A × B = Residual 1	Density × Block	$(p-1)(q-1)$	$\sigma_\varepsilon^2 + D_r \sigma_{\alpha\gamma\beta}^2 + r\sigma_{\alpha\beta}^2$	No test
C	Predators	$r-1$	$\sigma_\varepsilon^2 + D_q D_p \sigma_{\alpha\gamma\beta}^2 + D_q p\sigma_{\alpha\gamma}^2 + pqr\sigma_\gamma^2$	$\frac{MS_C}{MS_{CB}}$
C × B = Residual 2	Predators × Block	$(r-1)(q-1)$	$\sigma_\varepsilon^2 + D_p \sigma_{\alpha\gamma\beta}^2 + p\sigma_{\gamma\beta}^2$	No test
A × C	Density × Predators	$(p-1)(r-1)$	$\sigma_\varepsilon^2 + D_q \sigma_{\alpha\gamma\beta}^2 + q\sigma_{\alpha\gamma}^2$	$\frac{MS_{AC}}{MS_{ACB}}$
A × C × B = Residual 3	Density × Predators × Block	$(p-1)(r-1)(q-1)$	$\sigma_\varepsilon^2 + \sigma_{\alpha\gamma\beta}^2$	No test

Note:
There is only one replicate of each AC combination in each block (B). Components for fixed and random factors in expected mean squares are represented as "variances" – see Box 9.8.

Table 10.7 ANOVA for factorial randomized complete block design from Table 10.6 assuming that all block by factor interactions (A × block, C × block, A × C × block) are zero and are pooled into residual

Source	Wagner & Wise (1996)	df	Expected mean square	Test (A, C fixed)
B = Block	Block	$q-1$	$\sigma_\varepsilon^2 + pr\sigma_\beta^2$	$\frac{MS_{Block}}{MS_{Residual}}$
A	Density	$p-1$	$\sigma_\varepsilon^2 + qD_r\sigma_{\alpha\gamma}^2 + qr\sigma_\alpha^2$	$\frac{MS_A}{MS_{Residual}}$
C	Predators	$r-1$	$\sigma_\varepsilon^2 + qD_p\sigma_{\alpha\gamma}^2 + qp\sigma_\gamma^2$	$\frac{MS_C}{MS_{Residual}}$
A × C	Density × Predators	$(p-1)(r-1)$	$\sigma_\varepsilon^2 + q\sigma_{\alpha\gamma}^2$	$\frac{MS_{AC}}{MS_{Residual}}$
Residual	Residual	$(q-1)(pr-1)$	σ_ε^2	

Note:
Components for fixed and random factors in expected mean squares are represented as "variances" – see Box 9.8.

(Section 10.3.1). Second, the pooled residual term requires a restrictive omnibus sphericity condition (Kirk 1995), which also cannot easily be checked. We recommend that the non-additive model and separate error terms should be used.

Our earlier comments about using time as a blocking factor also apply to factorial randomized blocks. Factorial experiments are more costly than single factor experiments because of the larger number of combinations of the factors and it may not be possible to have enough experimental units to replicate such an experiment. Repeating the experiment through time and using time as a blocking variable is a useful option.

10.11.12 Incomplete block designs

Very occasionally, we may have an experimental design where we would like to block the treatments but the number of experimental units in each block is less than the number of treatments so we cannot have every treatment represented in each block. Under these circumstances, the trick is to allocate treatments to blocks so that relevant hypotheses can be tested although some interactions have to be assumed to be zero. The simplest arrangement is a balanced design where every pair of treatments occurs once (and only once) in one of the blocks. These designs can be arranged using randomized blocks or Latin squares and can also be unbalanced so that not every pair of treatments occurs in any block. The definitive reference is Cochran & Cox (1957) but Kirk (1995) and Mead (1988) also describe these designs.

Of course, some (including us) might argue that if there is such a mismatch between the available experimental units and number of treatment combinations, then reducing the number of treatments in the experiment is a more realistic solution. This is especially so in biology where treatment by block interactions are quite possible. The one exception might be where the design can be set up as a square arrangement with two blocking factors, as we will describe next.

10.11.13 Latin square designs

Sometimes we want to include two blocking factors in our design to further reduce the unexplained variation in our response variable. If the allocation of treatment levels to all combinations of blocking factors can be randomized, we could simply treat the combinations of the two blocking factors as levels of a single, combined, blocking factor and use the usual model for an RCB design. However, if we are willing to restrict the number of levels of each of the two blocking factors to be

A B C
B C A
C A B

C B A
A C B
B A C

B C A
A B C
C A B

A B C D
B C D A
C D A B
D A B C

B A D C
C D A B
D C B A
A B C D

D B C A
A D B C
C A D B
B C A D

Figure 10.7 Three possible random arrangements for 3 × 3 (three treatments: A, B, C) and 4 × 4 (four treatments: A, B, C, D) Latin squares.

the same as the number of treatment levels, we can also use a Latin square design. As the name suggests, Latin squares consider the experimental design as a square with equal numbers of rows and columns. One blocking factor is allocated to rows and the other to columns and there is a single experimental unit for each combination of row and column, i.e. cell. Latin square designs can be 2 × 2, 3 × 3, 4 × 4, etc. Treatments are allocated randomly to cells, with the restriction that each treatment is represented once in each row and in each column. By definition, the number of levels of factor A must be the same as the number of rows and the number of columns. Latin square designs are basically an extreme example of an incomplete block design, where the number of treatments represented in each block (row–column combination) is one!

There are many possible random arrangements of allocating treatments to cells in Latin square designs (Figure 10.7). For example, there are 12 possible arrangements for a 3 × 3 square and 576 arrangements for a 4 × 4 square. For a particular experiment, we simply select at random one of the possible arrangements of the appropriate size. Statistical software often include modules for the design of experiments that generate Latin square arrangements.

Traditionally, Latin square designs were used when the rows and columns represented a physical spatial arrangement of experimental units in the field. For example, Golden & Crist (1999) examined the effects of habitat fragmentation on old-field canopy insects using a 120 × 150 m field

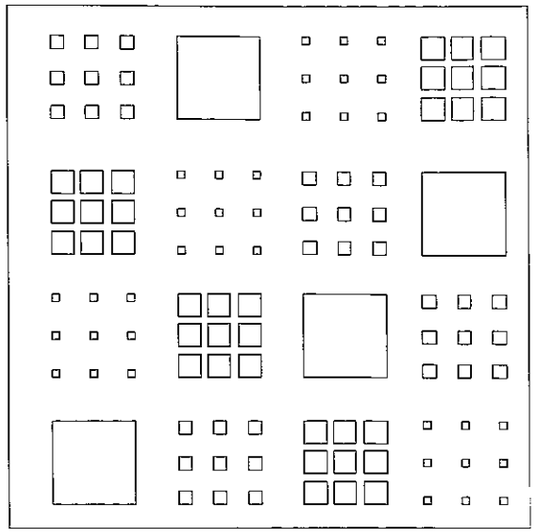

Figure 10.8 Layout of 4 × 4 Latin square experimental design from Golden & Crist (1999), showing four levels of fragmentation arranged in four rows and four columns.

(comprising goldenrod and wild carrot as dominant flora) in Ohio. They had four treatments (levels of factor A) set up by mowing: unfragmented, slightly fragmented (3 m^2 subplots separated by 2 m mown strips), moderately fragmented (2 m^2 subplots separated by 3.5 m mown strips) and heavily fragmented (1 m^2 subplots separated by 5 m mown strips). They created 16 plots (each 13 × 13 m) in four rows and four columns and allocated treatments in a four by four Latin square design, i.e. each treatment was represented once in each row and each column (Figure 10.8). Basically, this design is blocking treatments (fragmentation) against two blocking factors, rows and columns.

Latin square designs can also be used when the blocking factors do not really represent physical rows and columns. For example, Cochran & Cox (1957) describe an experiment where rows are five weeks, columns are the five days of the week, and each of five treatments was allocated to each combination of week and day in the usual manner.

Consider a Latin square design with factor A ($i = 1$ to p) being treatments, factor B ($j = 1$ to p) being rows and factor C ($k = 1$ to p) being columns. Each observation is y_{ijk} (the value in each cell), the marginal treatment means pooling rows and

columns are $\bar{y}_i$, the marginal row means pooling treatments and columns are $\bar{y}_j$ and the marginal column means pooling treatments and rows are $\bar{y}_k$.

The linear model used for a Latin square design is:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijk} \quad (10.14)$$

From Golden & Crist (1999):

$$(\text{species richness of insects})_{ijk} = \mu + (\text{fragmentation})_i + (\text{rows})_j + (\text{columns})_k + \varepsilon_{ijk} \quad (10.15)$$

In models 10.14 and 10.15 we find the following.

μ is the overall (constant) population mean, e.g. the overall mean number of insect species per leaf for all combinations of fragmentation treatment, row and column (i.e. all cells).

If factor A is fixed, α_i is effect of ith level of factor A ($\mu_i - \mu$) pooling over rows and columns, e.g. the effect of fragmentation on the number of species of insects, pooling rows and columns. If factor A 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 factor A that could have been used.

If rows are fixed, β_j is the effect of the jth row ($\mu_j - \mu$) pooling over levels of factor A and columns, e.g. the effect of the different rows on the number of species of insects, pooling fragmentation treatments and columns. If rows are random, β_j represents a random variable with a mean of zero and a variance of σ_β^2, measuring the variance in mean values of the response variable across all the possible rows that could have been used.

If columns are fixed, γ_k is the effect of the kth column ($\mu_k - \mu$) pooling over levels of factor A and rows, e.g. the effect of the different columns on the number of species of insects, pooling fragmentation treatments and rows. If columns are random, γ_k represents a random variable with a mean of zero and a variance of σ_γ^2, measuring the variance in mean values of the response variable across all the possible columns that could have been used.

ε_{ijk} is random or unexplained error associated with the observation at each combination of the ith level of factor A and jth row and kth column. For example, this measures the random error associated with the number of species of insects in each combination of fragmentation treatment, row and column. These error terms are assumed to be normally distributed in each cell, with a mean of zero ($E(\varepsilon_{ijk}) = 0$) and a variance of σ_ε^2.

Note that model 10.14 is an additive model with no interaction terms. The total number of experimental units is simply the total number of row and column combinations (p^2) with a single level of factor A allocated to each combination. With a Latin square design, it is not possible to estimate any interaction terms and therefore we cannot fit a non-additive model.

The ANOVA from fitting model 10.14 is presented in Table 10.8. The SS for factor A, rows and columns are calculated from the respective marginal means as usual. The $SS_{Residual}$ is simply the difference between these SS and SS_{Total}. With only one observation per cell in Latin square designs, we have no real estimate of σ_ε^2 unless we assume that all interactions between A, rows and columns are zero. Sometimes, the $SS_{Residual}$ is termed $SS_{Remainder}$ (Neter *et al.* 1996). The test for factor A simply uses the $MS_{Residual}$ as the denominator.

Factor A would usually be fixed in most biological applications. If we have a true physical Latin square where the rows and columns are spatial arrangements within that square, then they may be considered fixed because it is difficult to imagine from what populations of rows and columns they could be a random sample. If rows and columns are not spatial arrangements within a real square, then either might be considered random. The F test of factor A is the same no matter what combination of fixed and random factors we have in a Latin square design, although the H_0 and its interpretation will be different.

Latin square designs are quite restrictive in their application. They require that the number of levels of factor A equal the number of levels of the two blocking factors, rows and columns, although Mead (1988) describes alternative rectangular designs where only the number of rows or

Table 10.8 ANOVA table for standard Latin square design, where the number of levels of factor A (p) equals the number of rows (p) equals the number of columns (p). Factor A, rows and columns are fixed although their components in expected mean squares are represented as "variances"

Source	df	MS	Expected mean square	F	P
Factor A (fragmentation)	$(p-1)$	$\frac{p\sum_{i=1}^{p}(\bar{y}_i - \bar{y})^2}{p-1}$	$\sigma_\varepsilon^2 + p\sigma_\alpha^2$	5.13	0.043
		104.56			
Row	$(p-1)$	$\frac{p\sum_{j=1}^{p}(\bar{y}_j - \bar{y})^2}{p-1}$	$\sigma_\varepsilon^2 + p\sigma_\beta^2$	3.69	0.081
		75.22			
Column	$(p-1)$	$\frac{p\sum_{k=1}^{p}(\bar{y}_k - \bar{y})^2}{p-1}$	$\sigma_\varepsilon^2 + p\sigma_\gamma^2$	11.46	0.007
		233.73			
Residual	$(p-1)(p-2)$	$\frac{\sum_{j=1}^{q}\sum_{k=1}^{q}(y_{ijk} - \bar{y}_j - \bar{y}_k - \bar{y}_i + 2\bar{y})^2}{(p-1)(p-2)}$	σ_ε^2		
		20.40			
Total	(p^2-1)	15			

Note:
The example is from Golden & Crist (1999), July species richness data only.

columns matches the number of treatments. Additionally, the number of df for the residual is often small. For example, a 3×3 design will have only two df for the residual and 4×4 design will have only six df. In these circumstances, we might wish to also replicate at the level of squares so we have multiple Latin squares (Mead 1988). Probably the most serious restriction on the application of Latin square designs in biology is that there should be no interactions between treatments, rows or columns. While we can use Tukey's test for non-additivity (Section 10.3.2; Box 10.5) to check for some forms of interaction (Kirk 1995), it is difficult to imagine that, in field experiments, treatments would not interact with spatial rows or columns. If there are interactions, then the test of factor A is biased in a messy way (Kirk 1995). Finally, like all ANOVAs based on unreplicated factorial models, missing values cause real difficulties for Latin square analyses and our comments in Section 10.8 apply.

Latin square designs can become more complex than the standard design described here. For example, Graeco-Latin square designs allow for three blocking factors by superimposing two standard Latin squares (Cochran & Cox 1957, Kirk 1995, Mead 1988). The restrictions discussed above for standard Latin squares apply even more so for these complex extensions.

10.11.4 Crossover designs

An experimental design that combines attributes of Latin squares and repeated measures designs is the crossover design, often used in experiments that apply multiple treatments to individual organisms. In its simplest form, the crossover design can be considered as a Latin square where subjects are one blocking factor (e.g. rows) and time periods are a second blocking factor (e.g. columns) and treatments are applied to each combination of subject and period using one of the Latin square randomizations. Consider the study of Feinsinger *et al.* (1991) who examined competition between three species of forest understory plants in Central America. They set up an experiment to examine the effects of four treatments (relative densities of one species, either *Besleria* or *Palicourea*, and a second species *Cephaelia*: 10:10, 90:10, 10:90, 50:50) on response variables such as rate of hummingbird probes per flower or number of pollen tubes per style or number of seeds matured per flower. They had four time periods (either four or six days depending on the species) and used four focal plants (of either *Besleria* or *Palicourea*), which were the subjects, with a Latin square design as illustrated in Table 10.9. Actually, their experiment was more complicated because they replicated each square at three separate spatial blocks, but their basic unit was a single block (or square).

One of the characteristics of crossover designs is that different subjects receive the treatments in a different sequence, hence the value of the Latin square approach where each subject receives each treatment once but in a different order. So the effect of subjects (e.g. focal plants) in crossover designs is also an effect of sequence of treatments. Under some patterns of sequences across subjects, we may be able to separate out the effects of sequence from what are termed carryover effects. These are interaction effects between period and treatment and represent the effects of a preceding treatment independent of sequence. The pattern of treatment allocations must be a Latin square where every treatment follows or precedes every other treatment the same number of times because we can then measure carryover effects for all pairs of treatments without confounding with sequence. Not all randomization patterns for allocation of treatments to squares do this, but the pattern used by Feinsinger *et al.* (1991) did. For other patterns, sequence and carryover effects are confounded and cannot be separated. Note that when there are only two treatments (and two periods), the sequence and carryover effects are by definition the same. Really only simple carryover effects from the preceding treatment can be detected, rather than carryover effects from the preceding two or more treatments, unless we have a very large design.

Some details of the analysis of crossover designs can be found in experimental design texts like Cochran & Cox (1957), Crowder & Hand (1990), Mead (1988), Neter *et al.* (1996) and Yandell (1997), with a standard reference being Ratkowsky *et al.* (1993). We don't provide details on calculating the SS but the basic analysis from Feinsinger *et al.* (1991) is presented in Table 10.9. Basically the

Table 10.9 (a) Design of the crossover experiment from Feinsinger *et al.* (1991). Relative density treatments are indicated as A, B, C and D. Each square was replicated in three spatial blocks. (b) Analysis of crossover experiment from Feinsinger *et al.* (1991)

(a)

	Focal plant			
Time period	I	II	III	IV
1	A	B	C	D
2	B	D	A	C
3	C	A	D	B
4	D	C	B	A

(b)

Source	Single block (i.e. square) df	Replicated blocks (i.e. squares) df
Blocks		$(q-1)=2$
Focal plants, i.e. sequence (within blocks)	$(p-1)=3$	$(q\,(p-1))=9$
Periods	$(p-1)=3$	$(p-1)=3$
Periods × Blocks		$(p-1)(q-1)=6$
Treatments, i.e. relative density	$(p-1)=3$	$(p-1)=3$
Treatments × Blocks		$(p-1)(q-1)=6$
Carryover	$(p-1)=3$	$(p-1)=3$
Residual	rest = 3	rest = 15
Total	$(p^2-1)=15$	$(qp^2-1)=47$

Note:
First df column is from analysis of single block (square), second df is full analysis from three replicate blocks. There are $p=4$ focal plants, $p=4$ periods and $p=4$ treatments in each block (square) and $q=3$ blocks.

SS_{Total} is partitioned into $SS_{Treatments}$, SS_{Period} and $SS_{Sequence}$ based on marginal means, with the remainder forming the residual. This is the equivalent analysis from a Latin square design (Table 10.8). Feinsinger *et al.* (1991) could also measure carryover effects as a separate source of variation from the residual because their pattern of allocation of treatments to period and subject combinations had every treatment followed by every other treatment once. The number of carryover effects is the same as the number of treatments, as they are measuring the effect of each treatment on the one in the following period. If the rest period between treatments within a subject is long enough, there should be no carryover effects and Feinsinger *et al.* (1991) did not find any significant carryover effects in their study. Note that the correlations between repeated measures on the same subject, that require special consideration in the analyses of RM designs (Section 10.4.2), are assumed to be incorporated into the carryover effect (Yandell 1997).

Feinsinger *et al.* (1991) also replicated their basic Latin square in three spatial blocks, so their full design was a replicated Latin square (Yandell 1997) and the analysis included the block effect and

interactions between blocks and treatments, periods and sequences. Because the squares are replicated spatially (blocks), then periods are crossed with squares. If the squares are replicated through time, then the periods would be different for each square and periods would be nested within square (or block). In these analyses, all terms are tested against residual unless there are replicate subjects for each sequence, e.g. replicate focal plants for each sequence of treatments. Then there would also be a subjects within-sequence term that would be used for testing the sequence effect. In the example from Feinsinger *et al.* (1991), there was only one subject (focal plant) per sequence so there was no subject within-sequence term.

The limitations of these designs are the same as Latin square designs, primarily the assumed lack of interactions between treatments, periods and subjects and the few df for the residual, especially when carryover effects are separated out as a source of variation. Also, there are the usual difficulties of handling missing observations and the requirement that the number of treatments needs to match the number of subjects or periods. These designs are most commonly used in research on the responses of animals to different treatments where the number of animals is very restricted and both repeated measures on animals and through multiple time periods are needed.

10.12 Generalized randomized block designs

As we have emphasized, RCB designs are simply analyzed as unreplicated factorial ANOVAs. If replicates are possible within each combination of block and treatment, then we have a generalized randomized block design (GRB) whose advantages over the usual randomized block design include:

1. no need for any assumption of additivity,
2. separation of interaction effects from residual which may result in smaller $MS_{Residual}$ and more powerful test of treatments (Potvin 1993), and
3. better handling of missing values.

A GRB design that includes replicate experimental units for each treatment within each block is analyzed with a standard two factor linear model as described in Chapter 9 with a test for the factor A by block interaction. Note that randomization (random allocation of experimental units to treatments) is still restricted to n experimental units within each block, compared with a CR factorial design in which experimental units would be randomly allocated to each combination of the two factors. It is important that the "replicates" for a GRB design be at the appropriate scale, otherwise the usual factorial linear model is not applicable (Bergerud 1996). We must replicate the experimental units to which the levels of factor A are applied within each block, e.g. we must replicate leaves with and without domatia in each block in the example from Walter & O'Dowd (1992). If we simply subsample from each unreplicated treatment–block combination, e.g. we measure the size of individual mites in each combination of block (leaf pair) and treatment (with or without domatia), we can not use a two factor ANOVA model. We actually have a subsampled randomized block ANOVA where the analysis is as described in Table 10.10 for our fictitious modification of the Walter & O'Dowd (1992) experiment. Here, the non-existent true replicates for the two factor ANOVA model (replicate leaves for each treatment–block combination) are included in the ANOVA table to illustrate that the subsampled mites are not the appropriate replicates for testing any of the higher terms in the model (Bergerud 1996) – this is just a more complicated example of "pseudoreplication" (Hurlbert 1984; see also Chapter 7). Like Bergerud (1996), we suspect that many biologists mistake subsampling for true replication and would incorrectly analyze this design in Table 10.10 as a completely randomized two factor ANOVA.

10.13 RCB and RM designs and statistical software

Most statistical software distinguishes between RCB and RM designs in the way the data need to be coded. For an RCB design, each row in the data file represents an individual experimental unit, i.e. a treatment–block combination, and the data for the response variable are in a single column. The columns in the data file will be as follows.

Table 10.10 ANOVA table for a subsampled randomized block design, modifying Walter & O'Dowd (1992) so that 10 mites were sampled from each treatment–block combination, i.e. each single leaf for each treatment within each block

Source	df	Expected mean square
Treatment A	$p-1=1$	$\sigma_\varepsilon^2+\sigma^2_{\gamma(\alpha\beta)}+\sigma^2_{\alpha\beta}+\sigma_\alpha^2$
Block B	$q-1=13$	$\sigma_\varepsilon^2+\sigma^2_{\gamma(\alpha\beta)}+\sigma_\beta^2$
Treatment × block (A × B)	$(p-1)(q-1)=13$	$\sigma_\varepsilon^2+\sigma^2_{\gamma(\alpha\beta)}+\sigma^2_{\alpha\beta}$
Leaves (treatment and block) C(AB)	$pq(r-1)=0$	$\sigma_\varepsilon^2+\sigma^2_{\gamma(\alpha\beta)}$
Mites (leaves (treatment and block)) D(C(AB))	$pqr(n-1)=252$	σ_ε^2

Note:
For simplicity, expected mean squares are provided without multipliers and components for both fixed and random terms are indicated as variances – see Box 9.8. Note that the leaves nested within each treatment–block combination are included in the ANOVA table although their df equal zero because there is still only one replicate leaf for each treatment in each block.

Factor A	Block or subject	Response variable
Shaved	1	9
Unshaved	1	1
etc.		

The analysis then uses a linear model statement that includes a constant (grand mean), factor A and blocks (but no interaction). Output is standard ANOVA but usually with no adjusted univariate or multivariate tests.

For an RM design, each row represents a block or subject and the response variables for each treatment (e.g. time) are in separate columns. The columns in the data file will be as follows.

Block or subject	A_1 (Shaved)	A_2 (Control)
1	9	1
etc.		

The analysis uses software-specific repeated measures commands or menu options. Output is usually standard repeated measures ANOVA with unadjusted and adjusted univariate tests and multivariate tests, and also trend contrasts across treatment means.

Why the difference? It is probably due to most textbooks distinguishing the two types of designs, particularly the influence of statistical texts focusing on psychological and educational research, such as Winer *et al.* (1991). The point is that it doesn't matter which way you set the data file up, the analyses will be identical. It depends on whether you want an estimate of epsilon, measuring whether the variances and covariances meet the sphericity assumption and the extended output of adjusted univariate tests or multivariate tests – if so, use the repeated measures set-up.

10.14 General issues and hints for analysis

10.14.1 General issues

- Randomized complete block (RCB) and simple repeated measures (RM) designs are both analyzed using a linear model for a two factor ANOVA with n equals one in each cell.
- The test for factor A is $MS_A / MS_{Residual}$ whether there is an interaction between treatments and blocks/subjects or not.
- If treatment by block or subject interactions exist, then the power of the test for factor A is reduced and, if the interaction is strong, non-significant treatment effects are difficult to interpret.

- Blocks should normally be a random factor, otherwise there is no test for treatments unless we assume no treatment by block interaction.
- Violation of the sphericity assumption can seriously affect the univariate *F* tests and either adjusted univariate or multivariate tests of treatment effects should be used, especially in repeated measures situations.
- Factorial RCB designs are analyzed equivalently to factorial "within subjects" designs in repeated measures terminology, using the linear model for a three factor unreplicated factorial ANOVA. Each fixed main effect and interaction term should be tested against their interaction with block if blocks are random.

10.14.2 Hints for analysis

- For most statistical software, you should consider creating two data files, one coded for an unreplicated two factor crossed linear model analysis and one coded for classical repeated measures design. The basic ANOVA output will be the same, but other aspects of the output will differ and both contain useful information.
- Even though treatment by block interactions does not preclude assessment of treatment effects, it is worth running checks for interactions. If interactions are present, significant main effects interpretation is an effect of treatment over and above the interaction between treatment and blocks. It would also suggest that a generalized (i.e. replicated) RCB design should be considered if the experiment is repeated.
- Transforming lognormal data to logs or count data to a power (e.g. square or fourth root) can greatly improve additivity and should be used if the absence of treatment by block interactions is important for the analysis or interpretation.
- Cell mean plots are the simplest way of detecting treatment by block/subject interactions, although various residual plots can be also be helpful; a formal test for simple interactions is Tukey's single df test for additivity.
- Use separate denominators for F tests of contrasts between, or trends through, treatments.
- With missing values, either omit the block/subject with missing value if the number of blocks is large, or else estimate the missing value from marginal and overall means or use the model comparison approach as part of fitting the relevant linear models.

Chapter 11

Split-plot and repeated measures designs: partly nested analyses of variance

In Chapter 9, we described multifactor ANOVA models that can involve crossed or nested factors, or a combination of both, and in Chapter 10, we introduced designs that incorporate either blocks or repeated measures. One particular class of experimental designs with both crossed and nested factors, and either blocks or repeated measures, includes split-plot designs (from an agricultural origin), and repeated measures designs (from psychology). These designs can be complex but are particularly common in biological research, so we have devoted a chapter to their analysis. We will use the term partly nested or partly hierarchical for the linear model we fit with these designs, and the least ambiguous name for these designs might also be partly nested. One of the important messages from this chapter is that these repeated measures and split-plot designs are basically analyzed with the same linear model, something that is often unappreciated by biologists, although some textbooks do emphasize the equivalence in models (e.g. Kirk 1995, Mead 1988). In its simplest form, this design has three factors: A and C are crossed, and B is nested within A but crossed with C, although the possible extensions of this design are almost limitless.

11.1 Partly nested designs

11.1.1 Split-plot designs

Split-plot designs were originally used in agricultural experiments and represent a randomized complete block (RCB) design, with one or more factors applied to experimental units within each block. A second factor (or set of factors) is then applied to whole blocks, with replicate blocks for each level of this factor. Note that the terms blocks and plots are interchangeable in the context of these designs.

There are many examples of classical split-plot designs in the biological literature. First we will consider a fictitious extension of the RCB experiment we described in Chapter 10 from Walter & O'Dowd (1992), examining the role of domatia (small cavities on the leaf surface where mites can live) in determining the number of mites on leaves from species with domatia. They set up pairs of leaves (blocks) on a tree where one leaf in each pair was a control and the other leaf had its domatia removed. The treatment factor was applied to experimental units (leaves) within each block (leaf pair). If we now include additional plant species (those that have domatia), we now have a second factor applied at the scale of whole blocks, i.e. a block will be one or other of the species. This new experiment has blocks as the scale of replication for comparisons of species and leaves within blocks as the scale of replication for comparisons of treatments.

As another example, consider the experiment from Wissinger *et al.* (1996) who studied the effects of competition and water regime (hydroperiod) on the ecology of two species of larval caddisflies (*Asynarchus nigriculus* and *Limnephilus externus*) in ponds (Figure 11.1). The experiment was set up as a RCB design, with a block (i.e. plot) being a single pond, chosen for having some consistency in environmental conditions. Within each pond, they set