18

Statistical tools for analyzing data on behavioral ecology of insect parasitoids

Éric Wajnberg and Patsy Haccou

Abstract

Experiments performed by behavioral ecologists, both in the laboratory and in the field, are producing data that need to be accurately and properly analyzed by powerful statistical methods. Unfortunately, for most of the time and especially for experiments done on insect parasitoids:

1. these data are not following normal distributions (since counts, proportions, time duration, etc. are measured);
2. it is (nearly) impossible to use balanced or independent designs; and/or
3. there is pseudoreplication.

Thus, standard, well-established methods cannot be used and alternative statistical approaches have to be applied. This chapter gives a detailed overview on how the different problems mentioned above can be handled accurately.

18.1 Introduction

Nowadays, cheap computer devices and/or programs make it easy to collect large amounts of information from behavioral observations. As a consequence, behavioral ecologists, and especially those working on insect parasitoids, regularly need to handle a large variety of data. Results obtained from analyses of these data are usually confronted with ultimate predictions derived from theoretical models that tell us what animals should do to behave optimally. To arrive at the right conclusions, it is essential that the data are analyzed with correct and powerful statistical methods. Usually, only a small part of the collected data consists of values that are independent and have a Gaussian distribution. In this case, standard methods can be used, based on the so-called General Linear Model that includes most of the 'classical' methods that most readers will be familiar with, including linear regression, analysis of variance (ANOVA), and analysis of

covariance (ANCOVA). There are a large number of textbooks available that present these 'classical' methods (Zar 1999).

However, in an increasing number of cases, behavioral ecologists working on insect parasitoids also collect, and should thus also analyze, data that are either non-normally distributed – which is the case, for example, for time durations data, counts, or percentages – or that are non-independent or come from unbalanced experimental set-ups. Indeed, many of the chapters in this book deal with traits that are not normally distributed, such as time intervals (see also Chapter 8 by van Alphen and Bernstein, and Chapter 9 by Haccou and van Alphen) or sex ratios (see also Chapter 12 by Ode and Hardy). Further, especially when they are collecting data from fieldwork, scientists have to cope with pseudoreplication problems. In this chapter, we will present the basic theoretical framework and the statistical methods that can be used to handle such types of data correctly. Examples dealing with behavioral ecology of insect parasitoids will also be presented throughout.

18.2 An introduction to generalized linear models (GLMs)

All of us are in need of analyzing relations between observed values and a set of explanatory variables. Traditionally, this is done with ANOVA and regression methods. Such statistical tools have several limitations, like the assumption of a normal distribution for errors of measurement. Over the last few decades, several methods have been developed to deal with more general models. However, only a small number of behavioral ecologists working on insect parasitoids feel comfortable with, and have indeed applied these. Here we will provide an introduction to these methods that will hopefully overcome these problems and stimulate researchers in behavioral ecology of parasitoids to apply the methods of analysis most suited to their data.

18.2.1 The general framework

The framework we consider is as follows: We have a set of n observations $y_1, \ldots, y_n$, that are realizations of n random variables $Y_1, \ldots, Y_n$. These are usually called dependent variables. For the moment we assume that they are independently distributed. Further, we have a set of p explanatory variables, $x_1, \ldots, x_p$, with values associated to each of the n observations: $x_{11}, \ldots, x_{1p}; x_{21}, \ldots, x_{2p}; \ldots; x_{n1}; \ldots, x_{np}$. These variables are assumed to be known without error (in practice the methods also work well when these variables are random, with a much lower variance than the Y_i). They are also often called independent variables, or predictor variables. Note that the independent variables can be of several types. For instance binary (e.g. sex), nominal (e.g. colors), ordinal (e.g. sizes small, medium or large), or quantitative (e.g. weight in micrograms). For notational convenience, we will add a variable x_0 that equals one for every observation. This will allow us to formulate a null model of no effect in a notationally efficient way (see below).

We seek to relate the observations y to the set of explanatory variables x. This is done by assuming that the distribution of Y_i is somehow related to a predictor variable η_i, which is a linear combination of the explanatory variables:

$$\eta_i = \sum_{j=0}^{p} \beta_j x_{ij} \tag{18.1}$$

The coefficients β_j are estimated from the data. For instance, in a regression analysis, these are the regression coefficients. In an ANOVA, they represent the estimated effects of the factors. The way in which this so-called linear predictor affects the distribution of the Y_i is determined by a function that relates a parameter of this distribution to the value of η_i. In general, this parameter is the expectation of Y_i, and here we will only consider such cases. If, for convenience, we denote the expectation by μ_i, then it is assumed that

$$\eta_i = g(\mu_i) \tag{18.2}$$

The function $g(\cdot)$ is called the link function (McCullagh & Nelder 1989). Note that in some texts the link function is the inverse of $g(\cdot)$.

In classical analyses, the Y_i follow normal distributions with an average value of μ_i and common variance σ^2, and:

$$\mu_i = \eta_i = \sum_{j=0}^{p} \beta_j x_{ij} \tag{18.3}$$

In this case, the link function is thus the identity function. The variance σ^2 is an example of a so-called nuisance parameter, which means that this parameter is not related to the question that we are examining and, ideally, we do not want it to affect the outcome of the analysis in any way.

The meaning of Equations (18.1) to (18.3) is clear for the situation in which the explanatory variables are all quantitative, a situation that corresponds to a simple regression analysis. However, in order to express the effects of qualitative variables in this way, explanatory variables have to be assigned a value. Binary variables can, for instance, be coded 0 or 1. Nominal or ordinal variables with k classes can be represented by a set of $k-1$ so-called dummy variables, each taking the value 0 or 1. Table 18.1 gives two possible ways of using such dummy variables to code a nominal trait with three classes: Red, Yellow, and Blue.

For non-quantitative variables, the effects of the different classes can be derived from the coefficients β_j. For method A in Table 18.1, for instance, the coefficients for the two dummy variables give the effects of class Blue and Yellow, respectively, relative to class Red. The effect of Blue relative to that of Yellow is given by the difference between the coefficients of the two variables. If method B is used instead, the coefficient of the first dummy variable gives the effect of Blue relative to Yellow, whereas the effect of Blue

Table 18.1 Two equivalent methods of encoding a three-category nominal variable with two binary variables.

	Method A		**Method B**	
	Var1	**Var2**	**Var1**	**Var2**
Red	0	0	0	0
Yellow	0	1	0	1
Blue	1	0	1	1

relative to Red is given by the sum of the coefficients of the two dummy variables. Thus, the two methods give the same results for the relative effects, as they should.

When there are only qualitative variables, the classical analysis is the ANOVA. When there are qualitative as well as quantitative variables we are in the context of an ANCOVA. Interactions between the effects of the predictor variables can be examined by adding new predictor variables to the model, that consist of combinations of other variables.

The objectives of analyses are to estimate the coefficients β_j from the data, to select a model by identifying those explanatory variables (and their interactions) that have significant effects, to test goodness-of-fit of the selected model, and finally to make predictions based on the model. Prediction can concern, for example, the outcome of an experiment if the range of values of one of the explanatory variables is changed.

18.2.2 The Gaussian case

Estimation

Here, the Y_i are assumed to be random variables following independent normal distributions $N(\mu_i,\sigma^2)$, where the μ_i are given by Equation (18.3). We will first consider the situation where the common variance σ^2 is known. The probability density functions of the Y_i are thus of the form:

$$f(y_i;\mu_i,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(y_i-\mu_i)^2}{\sigma^2}\right] \tag{18.4}$$

Thus, the log-likelihood of all observations equals

$$L(\mathbf{y};\boldsymbol{\mu},\sigma^2) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{\sigma^2} \tag{18.5}$$

Note that we denote vectors by bold face, for example, $\mathbf{y} = y_1, \ldots, y_n$. The estimates of the regression parameters β_j are those values that maximize this expression (Mood et al. 1974). As can easily be seen from Equation (18.5), this implies that they minimize the sum of squares:

$$\sum_{i=1}^{n}(y_i-\mu_i)^2 = \sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p}\beta_j x_{ij}\right)^2 \tag{18.6}$$

which is a well-known result (Mood et al. 1974). There are a lot of statistical software packages that contain pre-programmed algorithms for finding the least-squares (or, equivalently, the maximum likelihood) estimates of the β_j for different types of models, but we will not go into the details of those methods here.

When the number of observations n is sufficiently large, the maximum likelihood estimators $\hat{\beta}_j$ have asymptotically a multivariate normal distribution with expected values β_j under the hypothesis that the fitted model is true. This result holds in general for maximum likelihood estimators, regardless of the distribution of the Y_i. Most statistical computer programs provide the estimated variance/covariance matrix of this multivariate

distribution. Confidence intervals for the estimates can be based on this distribution, or alternatively on the fact that the distribution of

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' V(\hat{\boldsymbol{\beta}})^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \tag{18.7}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$ is the vector of regression coefficients, $\hat{\boldsymbol{\beta}}$ is the corresponding vector of its estimators, and $V(\hat{\boldsymbol{\beta}})^{-1}$ is the inverse of the variance/covariance matrix of the parameters, tends to a χ^2 distribution with $p + 1$ degrees of freedom (denoted by *df*). We illustrate this with an example for a situation with two estimated parameters (i.e. with $p = 1$), β_0 and β_1. The estimated values of respectively the parameters and their variance/covariance matrix are

$$\begin{aligned} &\hat{\beta}_0 = 0.9, \hat{\beta}_1 = 2.1, \\ &V(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} 1.2 & 0.3 \\ 0.3 & 0.6 \end{pmatrix}, \end{aligned} \tag{18.8}$$

with inverse:

$$V^{-1}(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} 0.95 & -0.48 \\ -0.48 & 1.90 \end{pmatrix} \tag{18.9}$$

For large sample sizes, we thus find that the following statistic has a χ^2 distribution with 2 *df*:

$$\begin{aligned} &(0.9 - \beta_0, 2.1 - \beta_1) \begin{pmatrix} 0.95 & -0.48 \\ -0.48 & 1.90 \end{pmatrix} \begin{pmatrix} 0.9 - \beta_0 \\ 2.1 - \beta_1 \end{pmatrix} \\ &= 7.3341 + 0.95\beta_0^2 + 0.306\beta_0 - 7.116\beta_1 - 0.96\beta_0\beta_1 + 1.9\beta_1^2. \end{aligned} \tag{18.10}$$

The 5% critical value of a χ^2 distribution with 2 *df* equals 5.99. Thus, the equation:

$$7.3341 + 0.95\beta_0^2 + 0.306\beta_0 - 7.116\beta_1 - 0.96\beta_0\beta_1 + 1.9\beta_1^2 = 5.99 \tag{18.11}$$

defines an ellipse that corresponds to the simultaneous 95% confidence interval of the two parameters.

Model selection: the deviance

The maximum possible value for the likelihood is when the model fully describes the observations, which is when the μ_i are equal to the y_i:

$$L(\mathbf{y};\mathbf{y},\sigma^2) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - y_i)^2}{\sigma^2} = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \tag{18.12}$$

Note that this model is fully uninformative, since it implies that we have a separate parameter for each observation. The maximum likelihood estimators of the μ_i are denoted by

$$\hat{\mu}_i = \sum_{j=0}^{p} \hat{\beta}_j x_{ij} \tag{18.13}$$

and the likelihood obtained with these estimators equals

$$L(\mathbf{y};\hat{\boldsymbol{\mu}},\sigma^2) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\sigma^2} \tag{18.14}$$

Two times the difference between the maximum and the observed likelihoods equals

$$2(L(\mathbf{y};\mathbf{y},\sigma^2) - L(\mathbf{y};\hat{\boldsymbol{\mu}},\sigma^2)) = \sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\sigma^2} \tag{18.15}$$

When the Y_i have a Gaussian distribution, this statistic has a χ^2 distribution with $n-(p+1)$ degrees of freedom, since there are $p+1$ estimated parameters. We will denote this distribution by $\chi^2_{n-(p+1)}$. In the context of GLMs, the right-hand side of Equation (18.15) is called the scaled deviance. In this case, the deviance itself equals

$$D(\mathbf{y},\boldsymbol{\mu}) = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 \tag{18.16}$$

and model selection, as well as goodness-of-fit procedures for GLMs, are based on these quantities. In particular, differences between (scaled) deviances are used to compare nested models. For instance, suppose we want to test whether a factor with q different levels has a significant effect (i.e. a one-way ANOVA). As explained above, such a factor is represented by $q-1$ binary variables. Thus, the corresponding model has q parameters (because of the additional parameter β_0), and the scaled deviance for this model will follow a χ^2_{n-q} distribution. The estimator of β_0 in the null model is $\bar{y}$, the mean value of the y_i, and its scaled deviance therefore equals

$$\frac{D(\mathbf{y},\boldsymbol{\mu}\mathbf{0})}{\sigma^2} = 2(L(\mathbf{y};\mathbf{y},\sigma^2) - (\mathbf{y};\bar{\mathbf{y}},\sigma^2)) = \sum_{i=1}^{n}\frac{(y_i - \bar{y})^2}{\sigma^2} \tag{18.17}$$

which follows a χ^2_{n-1} distribution. A test whether the examined factor has a significant effect is based on the difference between the two scaled deviances:

$$\frac{D(\mathbf{y},\boldsymbol{\mu}\mathbf{0})}{\sigma^2} - \frac{D(\mathbf{y},\boldsymbol{\mu})}{\sigma^2} = \sum_{i=1}^{n}\frac{(y_i - \bar{y})^2}{\sigma^2} - \sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\sigma^2} \tag{18.18}$$

This statistic has a χ^2 distribution with $(n-1)-(n-q) = q-1$ degrees of freedom under the null model, and thus the effect of the factor can be statistically tested.

Until now we have assumed that the value of the nuisance parameter σ^2 is known. When this is not true, it can be estimated by

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{18.19}$$

Substituting this in Equation (18.18) gives the test statistic:

$$\frac{D(\mathbf{y},\boldsymbol{\mu}\mathbf{0})}{\hat{\sigma}^2} - \frac{D(\mathbf{y},\boldsymbol{\mu})}{\hat{\sigma}^2} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{18.20}$$

For large n, this statistic is known to follow a χ^2_{q-1} distribution. An alternative is to use the following test statistic:

$$\frac{\left(\sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2\right)\frac{1}{q-1}}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{18.21}$$

which follows a $F_{(q-1,n-1)}$ distribution under the null-hypothesis of no effect. This distribution is exact, i.e. it also holds for small n, and so provides a better test for small sample sizes. Note that the right-hand side of Equation (18.19) equals $D(\mathbf{y},\boldsymbol{\mu}\mathbf{0})/(n-1)$, so we can write Equation (18.21) as

$$\frac{(D(\mathbf{y},\boldsymbol{\mu}\mathbf{0})-D(\mathbf{y},\boldsymbol{\mu}))\frac{1}{q-1}}{D(\mathbf{y},\boldsymbol{\mu}\mathbf{0})\frac{1}{n-1}} \tag{18.22}$$

This demonstrates that an ANOVA is equivalent to an 'analysis of deviance' for the model considered, i.e. the linear model with independent normally distributed error terms and an identity link function. The same can be shown for a regression analysis or multifactor ANCOVA, where tests are also based on ratios of sums of squares.

Note that tests of effects and concomitant model selection can also be based on the asymptotic distribution of the β_j (Section 18.2.2). Asymptotically, the two procedures are equivalent. For small n, however, they can lead to different conclusions.

The two-way layout: orthogonality of effects

The multifactor ANOVA has a large advantage that most other analyses do not have, namely so-called orthogonality of effects. This means that the outcomes of tests for a certain (combination of) variables(s) do not depend on whether or not other variables are included in the null-hypothesis model. For example, consider the two-way layout, with two factors f_1 and f_2. We consider four nested models:

1. M_0: the null-model where none of the factors are included;
2. M_1: the model with only f_1;
3. M_{12}: the model with both f_1 and f_2; and
4. M_{1*2}: the model with f_1, f_2 and an interaction effect between f_1 and f_2.

Suppose also that f_1 has four levels and f_2 has three. This means that there are six more predictor variables, taking 0/1 values: three for f_1 and two for f_2. Table 18.2 gives a possible way of coding all these predictor variables. The expressions for μ_i corresponding to the different models are summarized in column 2 of Table 18.3. We also introduce some notation (in column 3) to be able to distinguish the models in the subsequent equations. Note that the (estimators for the) values of the coefficients β_j will differ for the different models. The fourth column gives the expression for the sum of squares to be minimized to estimate the coefficients β_j. The deviance for the null model M_0 is given in Equation (18.17). The other deviances are calculated by substituting the estimates of the β_j in the

Table 18.2 One example of two ordinal variables that are encoded by means of binary variables. The number of binary variables needed for one ordinal variable equals the number of levels of the ordinal variable minus one. Further, to represent the interactions between the two factors we need six more variables: $x_6 = x_1 \times x_4$, $x_7 = x_2 \times x_4$, $x_8 = x_3 \times x_4$, $x_9 = x_1 \times x_5$, $x_{10} = x_2 \times x_5$, $x_{11} = x_3 \times x_5$. For example the interaction between the fourth level of f_1 and the second level of f_2 is represented by $x_6 = 1$, $x_7 = 1$, $x_8 = 1$, $x_9 = 0$, $x_{10} = 0$, $x_{11} = 0$.

Levels of f_1	Variables for f_1			Levels of f_2	Variables for f_2	
	x_1	x_2	x_3		x_4	x_5
1	0	0	0	1	0	0
2	1	0	0	2	1	0
3	1	1	0	3	1	1
4	1	1	1	–	–	–

Table 18.3 Overview of the different models used as examples for testing the effect of two factors on the average value of a Gaussian trait, their parameters, and corresponding degrees of freedom of their residual sums of squares. See text for a detailed explanation.

Model	μ_i	Notation	Minimization of	Degrees of freedom
M_0	$\beta_0 x_{i0} (= \beta_0)$	$\mu 0$	$\sum_{i=1}^{n} (y_i - \mu 0)^2$	$n - 1$
M_1	$\sum_{j=0}^{3} \beta_j x_{ij}$	$\mu 1_i$	$\sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{3} \beta_j x_{ij} \right)^2$	$n - 4$
M_{12}	$\sum_{j=0}^{5} \beta_j x_{ij}$	$\mu 12_i$	$\sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{5} \beta_j x_{ij} \right)^2$	$n - 6$
M_{1*2}	$\sum_{j=0}^{11} \beta_j x_{ij}$	$\mu 1*2_i$	$\sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{11} \beta_j x_{ij} \right)^2$	$n - 12$

expressions in column 4. Degrees of freedom corresponding to the deviances are equal to the number of observations minus the number of estimated parameters. These are given in column 5.

To test whether inclusion of an additional term in the model has a significant effect, we can use an *F*-test, just as in the one-way ANOVA, as explained above. For instance, to calculate the numerator of the *F*-statistic for testing model M_1 versus M_0, we divide the difference between the deviances of the two models by the proper degrees of freedom. With the notation for the deviances introduced before (Equation 18.16), this gives

$$\frac{(D(\mathbf{y},\boldsymbol{\mu}\mathbf{0}) - D(\mathbf{y},\boldsymbol{\mu}\mathbf{1}))\frac{1}{3}}{D(\mathbf{y},\boldsymbol{\mu}\mathbf{0})\frac{1}{n-1}} \tag{18.23}$$

which follows a $F_{(3,\,n-1)}$ distribution. The outcome of this test tells us whether factor f_1 has a significant effect or not. A subsequent test of M_{12} versus M_1 provides the same information about the other factor. Note, however, that there is no *a priori* reason to include the two factors in the model in this specific sequence. We might as well do it the other way round. This would lead to a sequence of models M_0, M_2, M_{12}, and M_{1*2} (where M_2 denotes the model with only factor 2 included). In this case, a test as to whether the effect of f_1 is significant is performed by using the differences between the deviances of models M_2 and M_{12}.

In classical multifactor ANOVAs, the two expressions turn out to be equal. Thus, the sequence in which different factors are included in the model has no effect on their significance. This is a pleasant feature of these analyses, but it is unfortunately not usually the case, even in classical analyses beyond the ANOVA. A well-known example where this no longer holds true is, for example, in multiple regression, where there are several quantitative variables. There, it does make a difference if different ways of successively including predictor variables in the models are being used. In GLMs, other than the classical case, the effects of different factors are also usually non-orthogonal. Therefore, it is recommended not to consider just one deviance table, but several, corresponding to different model sequences.

Model selection: some further considerations

Finally, it has to be noted that model selection is usually not just a question of blindly performing a sequence of tests. It also involves knowledge of the biology behind the data, and often plain common sense. To give an example, tests may show that none of the main effects (or only one) are significant, whereas their interaction is. However, it is not a good idea to use a model with only interactions and no main effects (or only one main effect), since such a model is usually difficult to interpret. Another example is that any function of an independent variable may be included in the models. For instance, we may choose $\log(x_j)$, or its square root, as one of the predictor variables. Such variables may even give significant results. That does not mean, however, that it always makes sense to include them. In general, we would advise the use of only well-interpretable transformations. For instance, a multiplicative effect rather than an additive one might sometimes be more realistic. This can be realized by using log-transformations of the original variables. If we denote the original variables by z_j, then $x_j = \log(z_j)$ will be used in the model. In the Gaussian case, i.e. with an identity link function, this means that in terms of the original variables:

$$\mu_i = \sum_{j=1}^{p} \beta_j x_{ij} = \sum_{j=1}^{p} \beta_j \log z_{ij} = \log\left(\prod_{j=1}^{p} z_{ij}^{\beta_j}\right) \tag{18.24}$$

Biological reasoning might also bring us to include predictor variables in the model, even though their effects are not significantly different from zero, simply because we know that they should have an effect. Not finding the effect can be due to the fact that, with the current number of observations, the power of the test is too small compared to its magnitude. The opposite may also occur: if the number of observations is very large,

effects may be significant, even though they are very small. In this case, it is advisable to consider the relevance of the effect as a decision factor for including it in the model.

Goodness of fit: residuals

Once a model has been selected, we wish to get an impression of how good it describes the current data set. In the Gaussian case, this is usually done by studying the residuals corresponding to differences between observed values and those predicted by the model:

$$r_i = y_i - \hat{\mu}_i \tag{18.25}$$

If the fitted model is accurate, these residuals should follow a normal distribution with variance equal to σ^2. The so-called Pearson residual uses the estimated variance to scale this to a standard normal distribution:

$$\tilde{r}_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\sigma}^2}} \tag{18.26}$$

Note that the deviance equals the residual error sum of squares (Equation 18.16). In general we can write:

$$D(y,\mu) = \sum_{i=1}^{n} D(y_i,\mu_i) \tag{18.27}$$

where $D(y_i,\mu_i)$ indicates the contribution of observation i to the total deviance. This has a positive value. For GLMs, we can thus define a 'deviance residual':

$$r_i = sign(y_i - \hat{\mu}_i)\sqrt{D(y_i,\mu_i)} \tag{18.28}$$

In the Gaussian case this is equal to Equation (18.25). Unfortunately, for non-Gaussian models the distribution of these residuals is not always known.

18.2.3 The non-Gaussian case

The models in the previous section are widely applied. Their applicability, however, is restricted in a number of ways. One of them is that the data are continuous quantities that can assume any real value and that follow normal distributions. In practice, we often have to deal with different types of data. GLMs extend the framework of the classical analyses to such situations.

Binary data

An obvious model for binary data is based on the binomial distribution. The observations Y_i correspond to the number of times a 'success' is observed in conjunction with the covariate values $x_{i1}, \ldots, x_{ip}$. They are assumed to follow independent binomial distributions with parameters m_i and p_i, so we have

$$\Pr(Y_i = y_i) = \binom{m_i}{y_i} p_i^{y_i}(1 - p_i)^{m_i - y_i},\ y_i = 0, 1, \ldots, m_i \tag{18.29}$$

The expectation μ_i equals $m_i \times p_i$, and thus we can write this expression as

$$\Pr(Y_i = y_i) = \binom{m_i}{y_i}\left(\frac{\mu_i}{m_i}\right)^{y_i}\left(1 - \frac{\mu_i}{m_i}\right)^{m_i - y_i}, \; y_i = 0, 1, \ldots, m_i \tag{18.30}$$

This leads to the log-likelihood:

$$L(\mathbf{y};\boldsymbol{\mu}) = \sum_{i=1}^{n} \log\binom{m_i}{y_i} + \sum_{i=1}^{n} y_i \log\left(\frac{\mu_i}{m_i}\right) + \sum_{i=1}^{n} (m_i - y_i)\log\left(1 - \frac{\mu_i}{m_i}\right) \tag{18.31}$$

There are several link functions that can be used to specify the relationship between the expectations and the explanatory variables. We will consider the logistic function:

$$\eta_i = \log\left(\frac{\mu_i}{m_i - \mu_i}\right) \tag{18.32}$$

where η_i is the linear predictor given in Equation (18.1). In terms of the expected values, this gives

$$\frac{\mu_i}{m_i - \mu_i} = e^{\eta_i} \Rightarrow \mu_i = m_i \frac{e^{\eta_i}}{1 + e^{\eta_i}} \tag{18.33}$$

When the explanatory variables are all quantitative, this analysis is called a logistic regression (McCullagh & Nelder 1989). Maximization of Equation (18.31), with these expressions substituted, gives the estimators of the β_j. The deviance equals

$$2(L(\mathbf{y};\mathbf{y}) - L(\mathbf{y};\hat{\boldsymbol{\mu}})) = 2\left(\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + \sum_{i=1}^{n} (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - \hat{\mu}_i}\right)\right) \tag{18.34}$$

It has to be noted that the distribution of this statistic is usually not well-approximated by a χ^2 distribution (McCullagh & Nelder 1989). For large n, however, the differences between the deviances for different models approximately follow χ^2 distributions, with the numbers of degrees of freedom equal to the differences between the numbers of parameters in the different models.

In some cases, however, the binomial distribution does not adequately describe the data, due to so-called overdispersion. Such a phenomenon corresponds to the case in which observed variances are larger than the binomial variance $mp(1-p)$. This can occur, for instance, when the observations are in fact mixtures of independent, non-identified binomially distributed variables with (slightly) different success probabilities. Underdispersion can also occur, but its mechanistic explanation is not totally clear. In both cases, we may use a model where

$$E(Y_i) = m_i p_i, \; Var(Y_i) = \sigma^2 m_i p_i (1 - p_i) \tag{18.35}$$

with σ^2 being an unknown nuisance parameter, which is then also estimated from the data. A nuisance parameter is a parameter whose precise value is not of interest, and conclusions

from a statistical analysis should preferably be independent of this value. The test statistics now follows approximately a χ^2 distribution multiplied by $\hat{\sigma}^2$. The asymptotic variance/covariance matrix of the estimators is also multiplied by this value.

Counts

Sometimes observations can take any discrete value larger than or equal to zero. Examples are the size of egg batches laid by a gregarious pro-ovigenic parasitoid, the number of host encounters within a certain time interval, the number of females attacking simultaneously a patch of hosts (see also Chapter 9 by Haccou and van Alphen), etc. In this case, a model based on the Poisson distribution might be appropriate:

$$\Pr(Y = y) = e^{-\mu}\frac{\mu^y}{y!},\ y = 0, 1, \ldots \tag{18.36}$$

with log-likelihood:

$$L(\mathbf{y};\boldsymbol{\mu}) = \sum_{i=1}^{n}(y_i \log \mu_i - \mu_i - \log y_i!) \tag{18.37}$$

The usual link function for this model is

$$\eta_i = \log\ \mu_i \Rightarrow \mu_i = e^{\eta_i} \tag{18.38}$$

The deviance equals

$$2(L(\mathbf{y};\mathbf{y}) - L(\mathbf{y};\hat{\boldsymbol{\mu}})) = 2\sum_{i=1}^{n}\left(y_i \log\frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)\right) = 2\sum_{i=1}^{n} y_i \log\frac{y_i}{\hat{\mu}_i} \tag{18.39}$$

and the model is called a log-linear model (McCullagh & Nelder 1989).

It is well-known that the expectation of a Poisson distribution equals it variance. As for the Binomial model, however, overdispersion may inflate the variance. When observations correspond to numbers of occurrences in time intervals, overdispersion may be caused by random variation of the length of the intervals. Another possibility is inter-individual variability. The model may be generalized to account for such possibilities, by assuming:

$$Var(Y_i) = \sigma^2 E(Y_i) \tag{18.40}$$

where the nuisance parameter σ^2 can also be estimated from the data. Effects on tests and distribution of the estimators are accounted for in the same way as in the Binomial model.

Nominal or ordinal observations

Sometimes, the observations are counts of numbers in certain classes. Measurements can be on a nominal scale, for example, different substrates of a host species, or an ordinal scale, for example, we might measure the size of daughters emerging from a host as small, medium, or large. We then might be interested in effects of explanatory variables on the

probability that observations fall into a certain class. For instance, the host size might affect the proportions of small- and medium-sized daughters, compared to large ones. In this case, the appropriate model for the observations is a multinomial distribution:

$$\Pr(Y_1 = y_1, \ldots Y_r = y_r) = \frac{m!}{\prod_{k=1}^{r} m_k!} \prod_{k=1}^{r} p_k^{y_k} \tag{18.41}$$

with

$$\sum_{k=1}^{r} m_k = m, \sum_{k=1}^{r} p_k = 1 \tag{18.42}$$

where y_k denotes the number in class k. The log-likelihood becomes

$$L(\mathbf{y};\mathbf{m},\mathbf{p}) = \sum_{i=1}^{n} \log\left(\frac{m!}{\prod_{k=1}^{r} m_k!}\right) + \sum_{i=1}^{n}\sum_{k=1}^{r} y_{i,k} \log p_{i,k} \tag{18.43}$$

If we denote the expectation of the number in class k for the i^{th} observation by $\mu_{i,k}$, we can write this as

$$L(\mathbf{y};\mathbf{m},\boldsymbol{\mu}) = \sum_{i=1}^{n} \log\left(\frac{m!}{\prod_{k=1}^{r} m_k!}\right) + \sum_{i=1}^{n}\sum_{k=1}^{r} y_{i,k} \log \frac{\mu_{i,k}}{m_k} \tag{18.44}$$

and the deviance equals

$$D(y,m) = 2\sum_{i=1}^{n}\sum_{k=1}^{r} y_{i,k} \log\left(\frac{y_{i,k}}{\hat{\mu}_{i,k}}\right) \tag{18.45}$$

The linear predictor now becomes a vector with elements:

$$\eta_k = \sum_{j=0}^{p} \beta_{j,k} x_{ij,k}, k = 1, \ldots, r-1 \tag{18.46}$$

where $x_{ij,k}$ denotes the value of covariate j for category k of the i^{th} observation and, analogous to the previous models, $x_{i0,k} = 1$ for all i and k. The length of η is the number of categories minus one, due to the restrictions in Equation (18.42). To take these restrictions into account, the link function contains the ratio of expectations, relative to a chosen reference category. Here, we will use category r for that purpose. A logistic link function gives

$$\log\left(\frac{\mu_{i,k}}{\mu_{i,r}}\right) = \eta_k \Rightarrow \frac{\mu_{i,k}}{\mu_{i,r}} = e^{\eta_k}. \tag{18.47}$$

When explanatory variables are only quantitative, this again gives a logistic regression (as in the situation with binary variables (Section 18.2.3 above)). It should be noted that the ratio can be written in many different ways:

$$\frac{\mu_{i,k}}{\mu_{i,r}} = \frac{p_{i,k}}{p_{i,r}} = \frac{p_{i,k}}{1 - \sum_{k=1}^{r-1} p_{i,r}} = \frac{m_i p_{i,k}}{m_i\left(1 - \sum_{k=1}^{r-1} p_{i,r}\right)} = \frac{\mu_{i,k}}{m_i - \sum_{k=1}^{r-1} \mu_{i,k}}. \tag{18.48}$$

From this, it can be seen that the model is the multi-category analog of the one used for binary data (Equation 18.32).

It can be seen from Equation (18.46) that the number of parameters in a full model equals $(r-1)(p+1)$. An alternative special case of the model, often used, is

$$\eta_k = \beta_{0,k} + \sum_{j=1}^{p} \beta_j x_{ij,k} \tag{18.49}$$

where it is assumed that the covariates have the same effects on the different categories. This model only has $r + p - 1$ parameters.

When the data are measured on an ordinal scale, models can also be formulated in terms of the cumulative probabilities:

$$\gamma_{i,k} = \sum_{l=1}^{k} p_{i,l}, k = 1, \ldots, r-1 \tag{18.50}$$

In this case, a logistic model would be, for instance:

$$\log\left(\frac{\gamma_{i,k}}{1-\gamma_{i,k}}\right) = \beta_{0,k} + \sum_{j=1}^{p} \beta_j x_{ij,k} \tag{18.51}$$

However, it has to be noted that, in such models, the parameters are constrained since the cumulative probabilities $\gamma_{i,k}$ must increase with k. It can be deduced from Equation (18.51) that this implies that

$$\beta_{0,1} \leq \ldots \leq \beta_{0,r-1} \tag{18.52}$$

and any method used for estimation will have to take such constraints into account.

As in previous models, overdispersion may occur here, due to the same causes as mentioned before, and can be dealt with in the same way.

Survival data

In many situations, we have to deal with data that are time intervals until occurrence of certain events, or time intervals between events. For instance, behavioral ecologists often

record and analyze time intervals until the onset or termination of certain behaviors. Another example is residence time in host patches (see also Chapter 8 by van Alphen and Bernstein). Normal distributions usually do not give an accurate description of such data. Further, the data need special treatment due to the occurrence of so-called censoring. Indeed, sometimes the event in question is not observed during the experiment. For instance, a parasitoid may not leave a patch of hosts before the end of the observation period. When this occurs, we can only say that the patch-leaving time would have been longer than the observation time. Such observations cannot be simply treated as missing values, since they do contain information about the patch-leaving time. Bressers et al. (1991) give several examples of analyses that lead to erroneous conclusions due to the wrong treatment of censors. Survival analysis methods are especially designed to deal with this problem. Many of the models used in such analyses are, in fact, GLMs.

To illustrate this, we first consider the exponential regression model. In this model the Y_i are assumed to be exponentially distributed, with probability density function:

$$f(y;\lambda) = \lambda e^{-\lambda y} \tag{18.53}$$

where λ denotes the so-called hazard rate. In the exponential model this is a constant, but in more general models the hazard rate may depend on the time y. The hazard rate $\lambda(y)$ is the probability per time unit that an event happens at time y, given that it has not yet occurred. This is an important parameter in survival analysis, and most models are formulated in terms of this parameter rather than the expectation of Y. The interpretation of the hazard rate depends on the biological problem that is examined. For instance, if the observations are times until a patch of hosts is left, the hazard rate is the patch leaving tendency. In the case of, for example, grooming bouts, the hazard rate can be the tendency to start (or stop) grooming.

In the exponential regression model it is assumed that

$$\lambda_i = e^{\eta_i} \tag{18.54}$$

where η_i is the linear predictor of Equation (18.1). The null model parameter e^{β_0} is called the baseline hazard, and is denoted by λ_0, so an alternative way to denote the previous expression (which is the usual notation in survival analysis literature) is

$$\lambda_i = \lambda_0 e^{\sum_{j=1}^{p} \beta_j x_{ij}} \tag{18.55}$$

The expectation of Y_i equals

$$\mu_i = \frac{1}{\lambda_i} \tag{18.56}$$

and thus, we find that, in this case:

$$\mu_i = e^{-\eta_i} \Rightarrow \eta_i = -\log \mu_i \tag{18.57}$$

Thus, the exponential regression model is a generalized linear model with a logarithmic link function.

The probability of observing a time interval larger than some fixed value y equals

$$\Pr(Y > y) = e^{-\lambda y} \tag{18.58}$$

Censoring is taken into account by including this probability in the log-likelihood. For instance, if $y_1, \ldots, y_k$ are the time intervals observed in uncensored cases and $y_{k+1}, \ldots, y_n$ are censor times of the other observations, the log-likelihood equals

$$L(\mathbf{y};\boldsymbol{\lambda}) = \sum_{i=1}^{k} \log \lambda_i - \sum_{i=1}^{n} \lambda_i y_i \tag{18.59}$$

and the deviance equals

$$2(L(\mathbf{y};\mathbf{y}) - L(\mathbf{y};\hat{\boldsymbol{\lambda}})) = 2\left(\sum_{i=1}^{k} \log \frac{\hat{\lambda}_i}{y_i} + \sum_{i=1}^{n} \hat{\lambda}_i y_i - n \right) \tag{18.60}$$

and the residuals:

$$y_i - \frac{1}{\hat{\lambda}_i} \tag{18.61}$$

are known to have an exponential distribution with parameter 1 under the hypothesis that the model is correct (Kalbfleisch & Prentice 2002). Thus, the fit of a model can be tested by means of goodness-of-fit tests for an exponential distribution, where residuals of censored observations are treated as censors (see Haccou & Meelis 1994, for a description of several such tests).

The exponential regression model is a special case of the so-called Cox regression model, which is also called the proportional hazards model. In the more general formulation, the baseline hazard rate can be any positive function of the time interval y. This model assumes the following relationship between hazard rate and explanatory variables:

$$\lambda_i(y) = \lambda_0(y) e^{\sum_{j=1}^{p} \beta_j x_{ij}} \tag{18.62}$$

Hence, the variables are assumed to have a multiplicative effect on the baseline hazard. For instance, suppose that x_{11} equals A and x_{21} equals B, and that the other explanatory variables are all equal, i.e. $x_{1j} = x_{2j} = x_j$ for $j = 2, \ldots p$, then the model states that the ratio of the hazard rates is constant:

$$\frac{\lambda_1(y)}{\lambda_2(y)} = \frac{\lambda_0(y) e^{\beta_1 A + \sum_{j=2}^{p} \beta_j x_j}}{\lambda_0(y) e^{\beta_1 B + \sum_{j=2}^{p} \beta_j x_j}} = e^{\beta_1 (A-B)} \tag{18.63}$$

This is called the proportionality assumption. This assumption can be tested for any of the explanatory variables by fitting a model that allows a different baseline hazard for the

parameter in question, and then testing whether the estimated baseline hazards are indeed proportional. For instance, to test this for an explanatory variable X_1, which can assume two values, A and B, we use the model:

$$\lambda_i(y;A) = \lambda_{0A}(y)e^{\sum_{j=2}^{p}\beta_j x_{ij}}, \lambda_i(y;B) = \lambda_{0B}(y)e^{\sum_{j=2}^{p}\beta_j x_{ij}} \quad (18.64)$$

This is an example of a so-called stratified model. Proportionality tests of the baseline hazards can be formal or simply graphical (Kleinbaum & Klein 2005).

The probability density function is related to the hazard rate in the following way:

$$f(y;\lambda(y)) = \lambda(y)e^{-\int_0^y \lambda(s)\mathrm{d}s} \quad (18.65)$$

and the probability that Y is larger than y equals

$$\Pr(Y > y) = e^{-\int_0^y \lambda(s)\mathrm{d}s} \quad (18.66)$$

With these expressions we can write the log-likelihood, which is then maximized to estimate the β_j (j = 1, . . . , p), as well as the baseline hazard as a function of time. There are many computer programs available to do this. Tests of models can be based on the deviances or on the asymptotic multinormal distribution of the $\hat{\beta}_j$, as before. It can be shown that, here too, the residuals have an exponential distribution with parameter 1 if the model is correct (Kalbfleisch & Prentice 2002), which can be used to derive goodness-of-fit tests.

There are several generalizations of this model. An important one is to allow time-dependence of explanatory variables. Then the x_{ij} become functions of y also. For example, Hemerik et al. (1993) examined the effect of the number of rejected hosts after the most recent oviposition on the patch-leaving tendency. This covariate changes in time during a patch visit and was, therefore, modeled as a time-dependent variable.

Another generalization is to allow for repeated events. For instance, a patch might be left and revisited several times. The successive times on the patch can, for instance, be considered as separate observations, and we can include an extra covariate that counts the previous number of patch visits. Additional covariates may be, for example, durations of previous visits, or the durations of time intervals spent off the patch (see Wajnberg 2006, for a detailed recent survey of all parameters that were taken into account in studies on patch time allocation in insect parasitoids).

Finally, more than one type of event might be considered. For instance, a wasp on a patch might perform several types of behavior like walk, groom, rest, or examine a host. A walking bout can thus be followed by three different types of event, and we might be interested in studying effects of covariates on the hazard rates of each of these different types. These are called cause-specific hazard rates, and they can be studied separately, by considering other events as censors. For instance, to study the behavioral tendency to stop walking and start grooming, walking intervals that are followed by resting or host encounters are treated as censored observations. Methods for analyzing complete continuous time records of behavior can be found in Haccou and Meelis (1994).

The proportional hazards model assumes a multiplicative relationship between the explanatory variables and the hazard rate. An alternative is to assume such an effect on the time until occurrence of an event *Y*. This leads to the so-called accelerated failure time model. In the case of a constant hazard rate (i.e. the exponential model described previously), the models are equivalent. In the general case, however, this is not true. Accelerated failure time models are more difficult to interpret than proportional hazard models, and as a consequence are much less popular. Therefore, we will not describe these models here, but refer to the textbooks by Kalbfleisch and Prentice (2002), and Kleinbaum and Klein (2005) for details.

18.3 Non-independent data

The methods discussed so far assume that measures are all independent, which means that each data point has been collected from a different subject to ensure that the sample size reflects all independent replicates. However, some experimental set-ups used in behavioral ecology are explicitly based on so-called repeated measure designs in which subject are measured several times. This is, for example, the case for experiments done to estimate the learning ability of parasitoid females in which the same individuals are tested before and after a series of repeated experiences with hosts (van Baaren & Boivin 1998). As a general rule, repeated measures are produced in so-called longitudinal studies in which measures are collected at successive times. However, repetition can also sometimes correspond to data collected at different locations in space. Concerning behavioral studies done on insect parasitoids, this is especially the case, for example, in experiments done on four-way olfactometers (Vet et al. 1983) in which the behavior of the same individuals is recorded and compared between the different odorized fields of the device.

Using a standard ANOVA in this case is not appropriate because the data violate the assumption of independence and the analysis will fail to take proper account of correlations between the repeated measures. This can easily be seen from the following example based on simulated data: A normally-distributed behavioral trait is measured at three different times on the same individuals that were previously submitted to a specific treatment or a control group. We want to test both the effect of the treatment and any significant change in average values between the three recording times. This is clearly a repeated measure design but let us try to analyze it with a standard two-way ANOVA (Section 18.2.2). For this, values for the 3 times for 20 individuals were randomly drawn from 3 normal distributions, all having an average value of 0.0 and a standard deviation of 1.0. The first 10 individuals were supposed to experience the treatment, while the 10 remaining ones did not. In order to accurately simulate the repeated design, values corresponding to the three simulated times were randomly drawn from distributions and were correlated according to the following correlation matrix:

$$\begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}. \tag{18.67}$$

Different values of ρ, ranging from 0.0 to 1.0, were used and for each of them the whole simulation design was replicated 500 times. Each replicate was then analyzed with

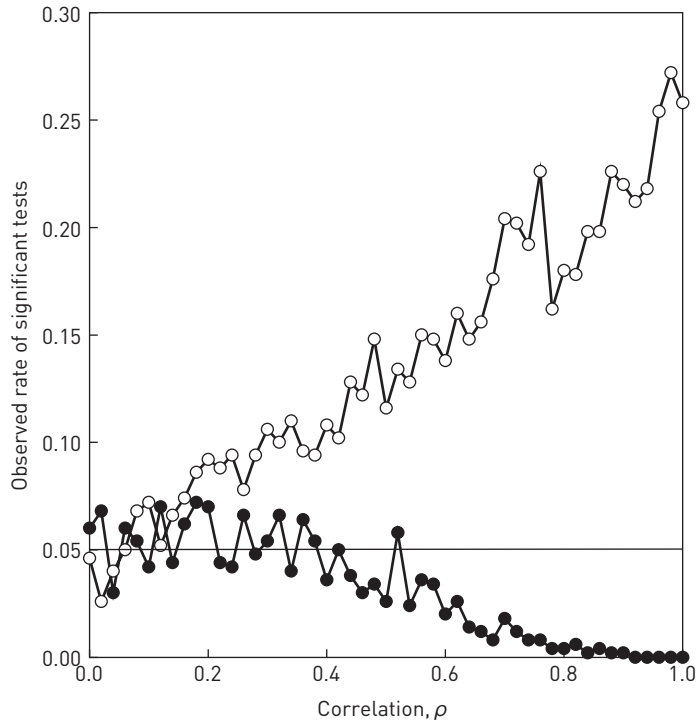

Fig. 18.1 Observed rate of significant tests for the treatment (open circles) and time (closed circles) effects after a two-way standard ANOVA on data simulated with different correlation values of the repeated design. Rates of significant tests are computed for a nominal level of 5%, which is indicated in this figure.

a standard two-way ANOVA to test both the treatment and the time effect, and the observed frequency of tests significant at a 5% level was computed in each case. Figure 18.1 gives the results obtained.

Since there were no differences between treatments and times in the simulated data, about 5% of the computed tests should be significant at a 5% level. As can be seen in Fig. 18.1, this appears to be the case only when the correlation ρ is close to 0.0, which corresponds to a situation of independent data. When the correlation ρ increases, however, the significance level of the test drops, which results in a larger number of falsely positive tests for the treatment effect. For the test of a time effect on such autocorrelated data, increasing values of the correlation ρ progressively lead to more conservative tests with a corresponding lack of power. At the extreme, when $\rho = 1.0$, the data remain unchanged over the course of time, and there cannot be a time effect.

Thus, standard ANOVA leads to wrong statistical conclusions, as it generally does when there are repeated measures, and so other methods should be used instead. One possible way to analyze data coming from a repeated design can be to consider the repeated measures as different variables (see Fig. 18.2 for a graphical meaning of this). Doing this, the correlation between the repeated measures among individuals is explicitly taken into account and standard multi-dimensional ANOVA (MANOVA) can then be used to test

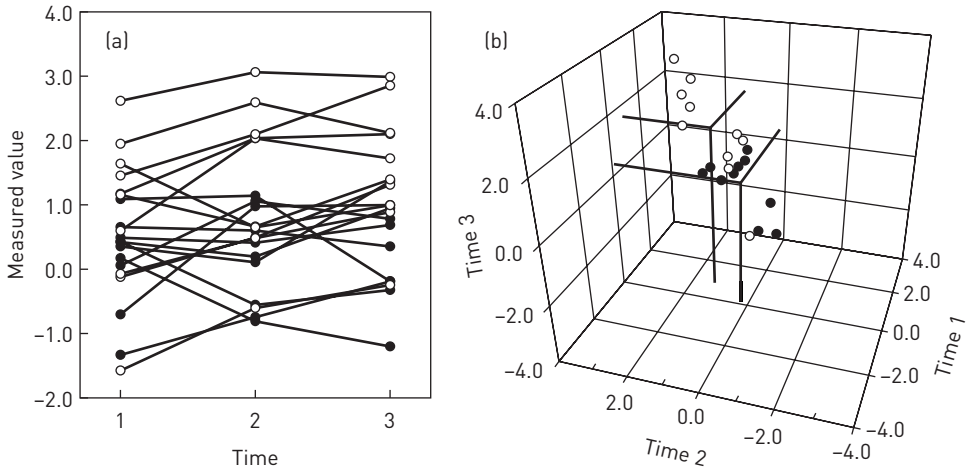

Fig. 18.2 Two graphical representations of the same data set showing how data coming from a repeated design can be considered as a multi- (here three-) dimensional data set. Data were simulated as in Fig. 18.1 (with $\rho = 0.75$), except that 1.0 was added to each value of individuals that were previously submitted to a specific treatment (open circles). (a) Values recorded at three different times are plotted as for a standard longitudinal study. (b) Data are plotted in a three-dimensional space and lines are indicating average values for individuals that were previously submitted or not to a specific treatment.

the different effects studied. Multivariate ANOVAs are simple generalizations of their standard univariate counterparts; the main difference is that inter- and intra-group variance-covariance matrices are used instead of the simple inter- and intra-group variances. Several tests can be used, for example, the Wilks' lambda (Johnson & Wichern 2002), to test, globally, the effect of factors that are not repeated, like the treatment effect in the simulated example above. The repeated factor (e.g. time) can also be tested, but after some data transformation. In the simulated example presented above, this will correspond to the test of the hypothesis that the three expected values μ_1, μ_2 and μ_3, corresponding to the three recording times, are equal. This is equivalent to testing simultaneously the two related hypothesis $\mu_1 - \mu_2 = 0$ and $\mu_2 - \mu_3 = 0$, which can be done with a Hotteling's T^2 test (Johnson & Wichern 2002). For this, each difference between the successive times, and their corresponding average values, should first be computed. Then, the following statistic:

$$\frac{(k - p + 1)}{kp} n \bar{Y}' \hat{V}^{-1} \bar{Y} \tag{18.68}$$

where n is the total number of individuals observed, p is the number of repeated measures (here 3), $\bar{Y}$ the mean average vector of the differences, $\hat{V}$ the estimated intra-group variance-covariance matrix, and k its corresponding *df* in a multivariate ANOVA done on the differences, is known to follow approximately an F-distribution with p and $k - p + 1$ *df*.

However, such an approach will be difficult to apply when the experimental set-up involves more than one factor, and their potential interactions, and/or when we are interested in testing interactions between the repeated factor (e.g. time) and others. Actually, a related method, so-called repeated measures ANOVA, is nowadays available in all software packages. It is based on the fact that, as opposed to data usually analyzed with multivariate analyses, repeated designs consist of collecting repeated measures of the same parameter. In repeated measures ANOVA, an individual is called a subject and the repeated factor is called a within-subjects factor, which represent different trials. Other factors are called between-subjects factors and are constituted of different groups. The method enables us to statistically test:

1. the within-subject main effect to know whether average values are changing between different trials;
2. between-subject main and interaction effects to estimate, globally, influences of the different corresponding factors; and also
3. within-subject-by-between-subject interaction effects to see whether changes among the different trials are influenced by any other factors.

For tests that involve only between-subjects effects, computations are simply based on simple ANOVA done on the sum of values obtained during the repeated trials divided by the square root of their number. Tests involving within-subjects effects are usually based on multivariate approaches similar to those described above (Davis 2003).

Repeated measures ANOVA carries the standard set of assumptions associated with an ordinary ANOVA, extended to the matrix case: multivariate normality for the within-subject factor, homogeneity of covariance matrices, and independence among groups for between-subject factors. Repeated measures ANOVA is robust to violations of the first two assumptions. Violations of independence among groups produce a non-normal distribution of the residuals, which results in invalid F tests. The most common violations of independence occur when either random selection or random assignment is not used. Some additional assumptions should also be verified, depending on the statistical test used to test the within-subject effect (see Davis 2003, for a thorough discussion of this).

Finally, as with fully independent data, there are now methods to analyze repeated designs in which the trait studied is not normally distributed. More accurately, the GLMs presented in Section 18.2 above can be extended to non-independent repeated designs by means of the so-called Generalized Estimating Equation of Liang and Zeger (1986), leading to methods for analyzing traits that follow a binomial (i.e. percentages) or a Poisson (i.e. counts) distribution (Hardin & Hilbe 2003).

18.4 Pseudoreplication

We have seen that experiments done on the behavioral ecology of insect parasitoids can sometimes produce non-independent data. This has been mainly presented to appear when each individual is measured several times (repeated measure designs), and we saw that specific statistical approaches are needed in this case (Section 18.3). For repeated measure designs, the experimental set-up has supposedly been built to intentionally produce

non-independent data. However, some experimental set-ups, that are not correctly designed, can sometimes also unintentionally and insidiously produce non-independent data that are then wrongly analyzed statistically. This is mainly, but not exclusively, observed in field experiments (Hurlbert 1984).

The problem was originally recognized by Hurlbert (1984), who analyzed 176 field experiments from 156 papers published in the ecological literature during 1960–1980. Among these, an alarming rate of 27% was guilty of so-called pseudoreplication. Considering only the 101 studies using statistical analyses, 48% were pseudoreplicated. Pseudoreplication is defined as the use of statistical methods to test for treatment effects with data from experiments where either treatments are simply not replicated (though samples may be) or experimental units are not statistically independent (Hurlbert 1984). Since the original paper of Hurlbert (1984), pseudoreplication has been widely reported in environmental, ecological, and behavioral studies (Steward-Oaten & Murdoch 1986, Searcy 1989, Hurlbert & White 1993, Heffner et al. 1996, Ramirez et al. 2000).

Potential problems might arise when treatments are spatially or temporally segregated. Ramirez et al. (2000) proposed a hypothetical, although classic, experimental example. The experiment aims at studying the ability of an insect parasitoid to be attracted from a plant attacked by one of its herbivorous hosts. Using an olfactometer, individual wasp females are offered a choice between two odorous areas: one receives volatiles from an attacked plant, the other volatiles from an unattacked plant. Fifty observations are performed and the attacked and unattacked plants are changed every five observations. The olfactometer is cleaned after each observation. With such an experimental set-up, using a simple *t*-test for paired data, for example, to compare walking parameters of the females in both areas over the 50 replicates, would be wrong for two reasons. The first one is due to the fact that volatiles coming for the attacked and unattacked plants are always released in the same areas of the olfactometer and some undetectable differences between these locations could generate differences in the recorded behavioral parameters that are not necessarily related to plant volatiles. This is an example of treatments that are spatially segregated, and this is the reason why it is usually proposed to rotate the olfactometer by 90 or 180° after each replicate. The other reason, that corresponds to a temporally segregation of treatments, is more insidious. It comes from the fact that plants are not changed after each observation but only after every fifth observation. Therefore, observations are not independent within groups. The ten groups of five observations, however, constitute legitimate replicates on which a simple *t*-test for paired data could be applied, and each is best represented by the mean value of the five observations performed in it (Ramirez et al. 2000). Another way to analyze the full data set is to use repeated measures ANOVA, as explained in Section 18.3 above (Ramirez et al. 2000).

Other potential problems are related to experiments that are actually interconnected. For example, an experiment performed on insects that were reared in four different climatic chambers will produce data that are not totally independent due to any (maybe even non-detectable) variation between the different climatic chambers. Finally, as we have seen in Section 18.3 above, pseudoreplication can also be generated when several observations are done repeatedly on the same individual (repeated designs).

Pseudoreplication is not truly a problem of experimental design itself. Rather, it is often the result of a combination of experimental design and an inappropriate statistical analysis (Ramirez et al. 2000). It is interesting to see that the problem is usually not understood and it is sometimes claimed by researches that all experiments, at the extreme, are

dependent since all of them are run on the same planet Earth. This is maybe the reason why poorly designed or incorrectly analyzed experimental work is literally flooding the ecological literature (Hurlbert 1984). Actually, pseudoreplication is the result of non-independent experimental units at the specific scale of analysis for the hypothesis being tested (Ramirez et al. 2000).

In any experimental field or laboratory work, it is known on first principles that, due to stochasticity, two or more objects are different whatever the trait measured. Then, if we increase the number of samples taken from each unit and statistically compare them (e.g. with a *t*-test or a simple ANOVA) with a nominal risk of, for example, 5%, the chance of finding a significant difference will increase with increase in the number of samples per unit. However, increasing the number of independent experimental units per treatment will not increase the chance of finding a significant difference under the null hypothesis. This has been proposed as a possible criterion for distinguishing pseudoreplication from true replication (Hurlbert 1984). Such a result can be more accurately understood using the following simulated example.

We wish to compare average values of a behavioral trait, related to wasp females foraging activity, between two wasp populations. The two populations are each constituted by several different families, each family originating from a mated female (i.e. isofemale lines). In the two populations, the trait studied is known to follow a normal distribution with an average value of 0.1 and a standard deviation of 1.0, but there are known differences in average values between the families. The first experimental design, using pseudoreplicates, is based on only two families, one taken in each population, and having an average value of 0.0 and 0.2, respectively. Samples of the same size are randomly drawn from each family. The second experimental design, using true replicates, is based on only one sample taken from each family and the same number of families is sampled in each population. For the two designs, the sampling protocol is repeated 500 times with different sample sizes and the two populations are compared in each case with a simple *t*-test. Figure 18.3 gives the observed frequency of tests significant at a 5% level, which was computed in both cases for different sample sizes. As expected, with a properly designed experiment based on true replicates, the probability of judging an effect as significant when there is no effect (i.e. an error of the first kind) remains constant and does not depend on sample sizes used. However, with a wrongly designed experiment using pseudoreplicates, the risk of wrongly declaring a non-existing effect as being significant clearly increases with sample size. The reason for this is that, in the later case, the null hypothesis tested is actually no longer that of no difference between the two populations compared, but that of no difference between the two families from which samples were taken.

Pseudoreplication is usually considered as an insidious bane. Even if some occurrences are clear-cut, like in the simulated example above, others can be more subtle and difficult to detect and require an accurate understanding of the system under study if the problem is to be avoided (Heffner et al. 1996). Common sense is usually needed along with biological knowledge, and sometimes intuition should be applied (Hurlbert 1984). Sometimes, replications are simply impossible or not desired, for example, if the cost in time and/or money of each of them is great. Experiments involving unreplicated treatments can sometimes be the only or best option (Hurlbert 1984). In such a case, erroneous use of statistical tests can lead to wrong conclusions, and it would be far better to recognize weaknesses in the experimental set-up used and to use descriptive statistics rather than formal tests of hypotheses.

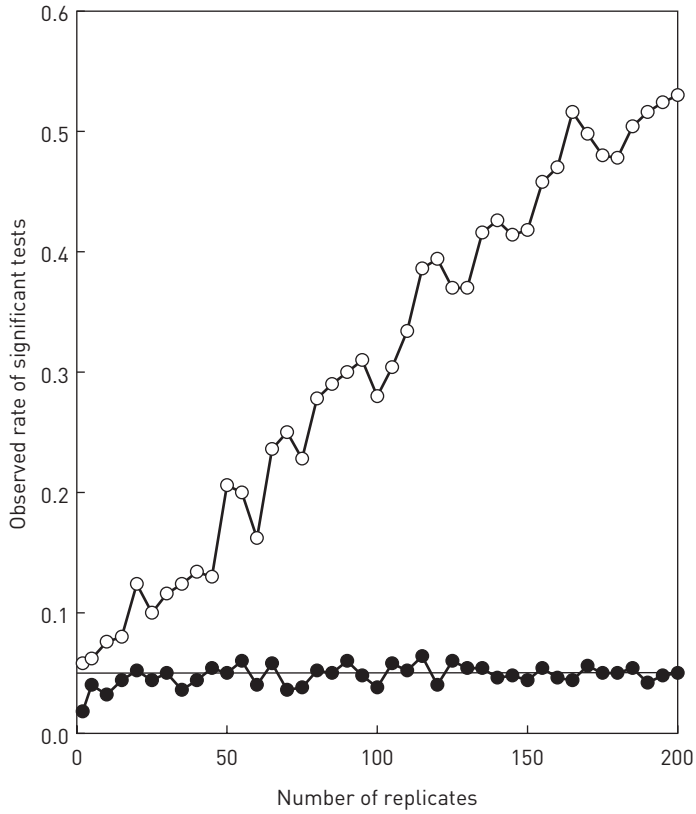

Fig. 18.3 Observed rate of significant tests comparing two wasp populations when true replicates (black circles) or pseudoreplicates (white circles) are used on simulated data with different sample sizes (see text). Rates of significant tests are computed for a nominal level of 5%, which is indicated in this figure.

18.5 Unbalanced set-ups

Many experiments to analyze the behavioral ecology of insect parasitoids are aimed at studying the effect of one or several independent qualitative factors on the average value of a behavioral trait. We already saw some examples of possible tested factors above. They usually correspond to different treatments whose effects have to be quantified and tested. When the effect of more that one factor is studied, their potential interactions are also of interest. In order to analyze the data obtained in such experiments, standard two-way (or more) ANOVAs (for normally-distributed traits) or GLMs (Sections 18.2.2 and 18.2.3) can be used. Usual statistical procedures are designed to handle so-called balanced data, which is data with equal numbers of observations for every combination of the different classification factors tested (Sokal & Rohlf 1981). However, even with the best of intentions, it is frequently impossible to produce such evenly balanced designs. For instance, in an experiment testing the effect of different treatments on wasp females on a behavioral feature of their offspring, the number of offspring per treated female is likely to

differ and will not be a parameter under experimental control. Moreover, even if an experiment is correctly designed, with a balanced number of replicates in each case, some individuals might accidentally die before being measured and the resulting data set will become unbalanced due to the existence of missing values.

Analyzing such unbalanced data is usually considered to be a complicated matter. One approach, if possible, is to avoid such a problem at the expense of losing some information. For example, if there are at least five replicates in every subclass in an analysis, but some of them have six of seven replicates, it would be possible to reduce the sample size of all subclasses to five to acquire a balanced design that may be analyzed with a standard ANOVA or a GLM. Of course, removal of any individuals from a subclass to equalize sample sizes must be done at random. Such a procedure will be legitimate only if the number of replicates in each subclass itself does not affect the trait under study. For example, in a gregarious parasitoid species, the behavior of all individuals that emerged from a single parasitized host might be influenced by their number. Further, reducing subclasses to a common sample size is usually not a good idea, especially if:

1. there is an important variation in subclass sample sizes;
2. the original data were scarce and/or expensive to obtain; or
3. the error (i.e. intra-subclass) variance is too important.

Another possibility, if the design is unbalanced due to missing values, is to estimate those values that are missing. If there are several replicates in each subclass, the missing values can be estimated by the observed average values of each corresponding subclass. If the behavioral trait under study is following a known distribution, missing data can also be approximated by simulated values randomly drawn from this distribution by using the average and variance of each corresponding subclass. Finally, if there is only one replicate per subclass, appropriate and usually easy to compute estimating methods are also available (Sokal & Rohlf 1981, Little & Rubin 1987). Once the missing values have been estimated, a standard ANOVA or a GLM can be computed, but the number of degrees of freedom of the error term has to be reduced by the number of values that have been estimated.

Estimating missing values is, however, not always feasible, especially if too many values are missing. Performing a two-way (or more) ANOVA or a GLM with unbalanced subclass sizes still remains possible, but statistical software packages are needed since computational procedures becomes considerably more complicated. In this case, a so-called type III ANOVA can be computed, which is designed to test the same hypotheses that would be tested if the data were balanced. Indeed, in this case, hypotheses and the associated tests are not functions of the number of replicates per treatment combination. On balanced data, such an ANOVA will give the same results as those obtained using standard ANOVA. However, results of such an analysis with unbalanced data should be interpreted cautiously (Shaw & Mitchell-Olds 1993). Probably the most important and obvious problem that can arise when analyzing unbalanced data by this method can be seen using the following example based on simulated values: A normally-distributed behavioral trait is measured on individuals belonging to four different wasp species and after four different treatments. In each species-treatment combination, three independent replicates were performed, so the full balanced design represents a total of $4 \times 4 \times 3 = 48$ replicates. All values were first drawn from a normal distribution with an average of 0.0

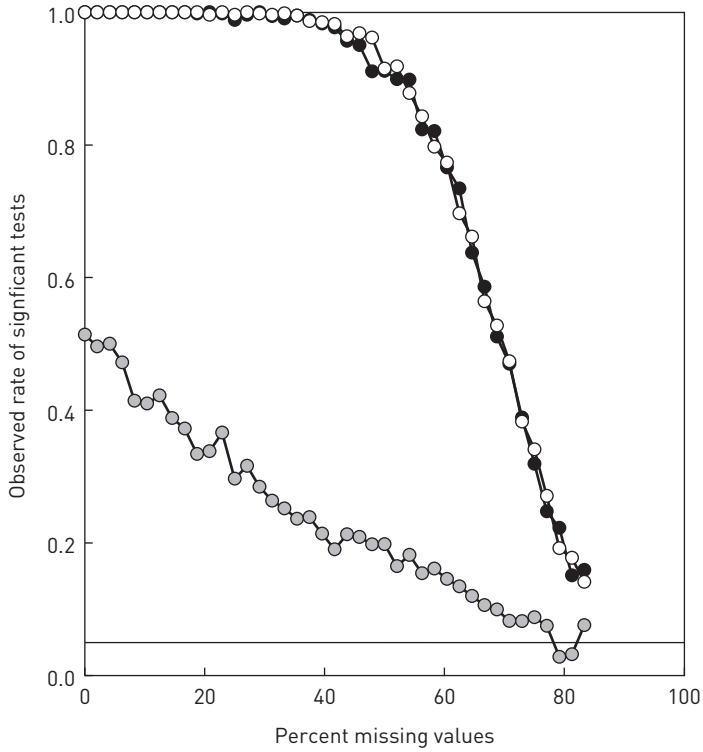

Fig. 18.4 Observed rate of significant tests for the species effect (white circles), the treatment effect (black circles), and their interaction (gray circles) after a two-way type III ANOVA on data simulated with different proportions on missing values (see text). Rates of significant tests are computed for a nominal level of 5%, which is indicated in this figure.

and a standard deviation of 1.0. In order to have significant effects for the two main factors and a significant interaction between them, $0.4 \times i \times j$ was added to each individuals belonging to the i^{th} species and that was treated with the j^{th} treatment (i and $j = 1, 2, 3, 4$), and the simulation design was replicated 500 times. By doing so, and using a nominal level of 5%, both the species and the treatment effect were always declared as being significant and their interaction was declared as being significant in 51.4% of the cases (Fig. 18.4).

The same computation was repeated each time with an increasing number of missing values that were uniformly distributed over the entire data set and the data were analyzed with a two-way type III ANOVA. The observed frequency of tests significant at a 5% level was computed in each case and results are shown in Fig. 18.4. As can be seen in this figure, an increase in the rate of missing values progressively leads to a reduction in the number of tests that are declared as significant. This is true for both the main factors tested and their interaction. For the main effects, the phenomenon seems to appear only when more than half of the data are missing but these factors were initially strongly significant (average p-values for the two factors with the full, balanced design were around 10^{-5} in both cases). On data simulated with less significant factors, the decrease in the power of the

corresponding tests would appear much earlier, as this can be seen for the interaction (average p-value for the interaction with the full, balanced design was around 0.112). Thus, the power of the test is reduced when data are unbalanced and it is, therefore, possible to overlook an effect (i.e. judge it as non-significant) when the effect truly exists. So, particular caution should be taken when interpreting failure to reject null hypotheses from unbalanced data (Shaw & Mitchell-Olds 1993).

18.6 Conclusion

Behavioral ecologists, and especially those working on insect parasitoids, regularly fall into standard traps when they analyze results of their experimental work. They indeed often have to deal with unbalanced or non-independent data, or handling pseudoreplications, sometimes even without knowing it. Further, they are in most cases collecting non-normally distributed data. In many cases, standard ANOVA (or regression methods) are applied, but arguments presented in this chapter show that this can lead to wrong conclusions about the effects that are tested. As we have shown, nowadays there is a whole arsenal of more rigorous and efficient methods available, which can be used in most of these cases. When no specific method is available (e.g. for pseudoreplicated data), we have tried to inform the reader about possible misinterpretation of the results obtained from wrongly designed statistical analyses. Thus, the objective of this chapter is to provide the reader with the basic knowledge for analyzing, in a more correct and efficient way, results collected from experimental field or laboratory works on the behavioral ecology of insect parasitoids.

Most of the statistical procedures presented in this chapter cannot be done by hand but a large number of statistical software packages are now available that incorporate these methods. Examples are:

1. SAS® (http://www.sas.com);
2. STATISTICA® (http://www.statsoft.com/);
3. SPSS® (http://www.spss.com/);
4. SYSTAT® (http://www.systat.com/);
5. Splus® (http://www.insightful.com/); or
6. R (http://www.r-project.org/).

The last one can be recommended because it is a free, efficient, and a simple-to-learn statistical computing and graphic language.

References

Bressers, W.M.A., Meelis, E., Haccou, P. and Kruk, M.R. (1991) When did it really start or stop: the impact of censored observations on ethological analysis of durations. *Behavioural Processes* **23**: 1–20.

Davis, C.S. (2003) *Statistical Methods for the Analysis of Repeated Measurements*. Springer, New York.

Haccou, P. and Meelis, E. (1994) *Statistical Analysis of Behavioural Data*. Oxford University Press, Oxford.

Hardin, J.W. and Hilbe, J.M. (2003) *Generalized Estimating Equations*. Chapman & Hall. CRC, Boca Raton, FL.

Heffner, R.A., Butler, M.J. and Reilly, C.K. (1996) Pseudoreplication revisited. *Ecology* **77**: 2558–62.
Hemerik, L., Driessen, G. and Haccou, P. (1993) The effects of intra-patch experiences on patch leaving tendency, search time and search efficiency of parasitoids of the species *Leptopilina clavipes*. *Journal of Animal Ecology* **62**: 33–44.
Hurlbert, S.H. (1984) Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* **54**: 187–211.
Hurlbert, S.H. and White, M.D. (1993) Experiments with freshwater invertebrate zooplanktivores: quality of statistical analyses. *Bulletin of Marine Science* **53**: 128–53.
Johnson, R.A. and Wichern, D.W. (2002) *Applied Multivariate Statistical Analysis*. Prentice Hall, Englewood Cliffs, NJ.
Kalbfleisch, J.D. and Prentice, R.L. (2002) *The Statistical Analysis of Failure Time Data*, 2nd edn. John Wiley & Sons, New York.
Kleinbaum, D.G. and Klein, M. (2005) *Survival Analysis – a Self-Learning Text*, 2nd edn. Springer, New York.
Liang, K.Y. and Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika* **73**: 13–22.
Little, R.J.A. and Rubin, D.B. (1987) *Statistical Analysis with Missing Data*. Wiley, New York.
McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*, 2nd edn. Chapman & Hall, London.
Mood, A.M., Graybill, F.A. and Boes, D.C. (1974) *Introduction to the Theory of Statistics*, 3rd edn. McGraw-Hill, Singapore.
Ramirez, C.C., Fuentes-Contreras, E., Rodriguez, L.C. and Niemeyer, H.M. (2000) Pseudoreplication and its frequency in olfactometric laboratory studies. *Journal of Chemical Ecology* **26**: 1423–31.
Searcy, W.A. (1989) Pseudoreplication, external validity and the design of playback experiments. *Animal Behaviour* **38**: 715–17.
Shaw, R.G. and Mitchell-Olds, T. (1993) ANOVA for unbalanced data: an overview. *Ecology* **74**: 1638–45.
Sokal, R.R. and Rohlf, F.J. (1981) *Biometry. The Principles and Practice of Statistics in Biological Research*, 2nd edn. Freeman, San Francisco.
Steward-Oaten, A. and Murdoch, W.W. (1986) Environmental impact assessment: 'pseudoreplication' in time. *Ecology* **67**: 929–40.
van Baaren, J. and Boivin, G. (1998) Learning affects host discrimination behavior in a parasitoid wasp. *Behavioral Ecology & Sociobiology* **42**: 9–16.
Vet, L.E.M., van Lenteren, J.C., Heymans, M. and Meelis, E. (1983) An airflow olfactometer for measuring olfactory responses of hymenopterous parasitoids and other small insects. *Physiological Entomology* **8**: 97–106.
Wajnberg, É. (2006) Time-allocation strategies in insect parasitoids: from ultimate predictions to proximate behavioural mechanisms. *Behavioral Ecology & Sociobiology* **60**: 589–611.
Zar, J.H. (1999) *Biostatistical Analysis*, 4th edn. Prentice Hall, New Jersey.