Chapter 3

Statistical analysis of sex ratios: an introduction

Kenneth Wilson & Ian C.W. Hardy

3.1 Summary

In this chapter we discuss how to make best use of sex ratio data. We identify three basic questions that such data can be used to answer: does the sex ratio differ from some theoretically expected *mean* value, does it differ from an expected *distribution* and is variation in sex ratio associated with some measured *explanatory terms*? Our main focus is on the latter question. We discuss analytical methods in order of 'sophistication', starting with *nonparametric* methods (which make few assumptions about underlying statistical distributions), then *classical parametric* methods (which assume that data conform to a normal distribution of deviations from a statistical model) and finally *generalized linear models* (*GLMs*). GLMs are semi-parametric methods that encompass models assuming a normal distribution but may also assume other distributions. This is an important advantage as sex ratio data are best expressed as proportions (sex ratio = males/(males + females)) and deviations are expected to conform to a *binomial* distribution. GLMs assuming binomial distributions are often termed *logistic regression models*. Distributions may not conform to the normal or binomial assumptions of classical parametric analyses or logistic GLMs, and we discuss how these problems can be overcome. The statistical approaches we discuss are illustrated with worked examples and case histories from recent sex ratio literature. We also perform simulations to evaluate the relative performances of non-parametric, classical parametric and logistic GLM analyses: GLMs win. A statistical analysis of the sex ratio literature published in 1994–2000 indicates that GLMs are currently being employed in only a small proportion (<30%) of sex ratio analyses and that the proportion does not appear to be increasing. Thus, this chapter serves in part as a manifesto for change, aimed at those who need to be persuaded that the GLM approach is worth learning and who need a short introduction to the subject.

3.2 Introduction

In this chapter, we present a guide to the statistical analysis of sex ratio data. Our aim is to present a brief introduction to statistical methods that will increase the accuracy and power of sex ratio analyses. As evolutionary ecologists (rather than statisticians), our emphasis here is on the practicalities of analysing sex ratio data and we aim to give an intuitive feel for the different methods, rather than to explore in depth their statistical basis. Readers interested in the formal proofs of the different methods we discuss should consult the following texts and original papers cited therein: Cox and Snell (1989), Hosmer and Lemeshow (1989), McCullagh and Nelder (1989), Collett (1991) and Crawley (1993, 2002). In the remainder of this introductory section we discuss the different ways in which sex ratio data can be expressed (section 3.2.1) and discuss the sorts of questions

that empiricists may want to ask about sex ratios (section 3.2.2). We then briefly outline possible analytical approaches used in answering these questions (section 3.3), assuming minimal statistical knowledge (Box 3.1 gives a refresher on statistical terminology). We introduce 'generalized linear models' (GLMs), a family of statistical tools, and we focus on logistic analyses, the sub-class of GLMs appropriate for analysing proportion data (section 3.4). We illustrate the relative merits of the different methods by means of a fictitious example (sections 3.3 & 3.4) and re-analyse some real datasets using this methodology (section 3.5). Although we employ a number of different nonGLM methods to analyse sex ratio data, we do this mainly to illustrate their lack of power and rigour and do not advocate their use except in exceptional circumstances. Finally, we illustrate the relative power of GLMs over alternative methods of analysis using a series of simple simulation studies (section 3.6).

Box 3.1 A brief introduction to statistical approaches

Statistics is all about differences and associations. Usually, we are asking questions such as 'Is *A* different from *B*?' or 'Are changes in *C* associated with changes in *D*'. Statistical tests allow us to assess whether differences or associations are **statistically significant** (i.e. whether observed patterns differ from those expected by chance alone). To do this, we generally formulate a **null hypothesis (H_0)**, i.e. we hypothesize that any observed difference or association is due to random effects. **Hypothesis testing** centres around either accepting or rejecting H_0. This decision is usually made by comparing the value of a **test statistic** with some predetermined **critical value** (which can be found in published tables, e.g. Rohlf & Sokal 1995) for a given **significance level**. Traditionally, this significance level is taken to be 0.05 or 5%. This means that one will reject the null hypothesis in favour of the **alternative hypothesis (H_1)** if the probability, *P*, that observed data could have arisen by chance alone is less than 5%. If it is (i.e. $P < 0.05$), then one may conclude that the difference is 'statistically significant'. Note that the choice of a particular significance level is an ultimately arbitrary convention that dichotomizes a continuum of probabilities. The lower the probability ($0.05 > 0.01 > 0.001$), the more sure one can be that the difference is not just random sampling error with no real underlying difference.

When we are hypothesis testing, we generally test rather general hypotheses (e.g. Does *A* differ from *B* or is there an association between *C* and *D*?), but at other times we may have *a priori* reasons for testing more specific hypotheses (e.g. Is *A* larger than *B* or is *C* positively correlated with *D*?). The former type of test is known as a **two-tailed test** and the latter a **one-tailed test**. This is because, in the first instance, we are testing for both positive *and* negative differences and correlations, whereas in the second we are just testing for positive (or negative) differences and associations. As a result, the critical value for rejecting the null hypothesis is increased and the associated *P* value is reduced (see chapter 7 in Sokal & Rohlf 1985).

If H_0 is rejected, H_1 is supported but not proven. Rejecting a correct H_0 is termed a **type I error** while failing to reject an incorrect H_0 is a **type II error**. The probability of committing a type I error is usually termed α and the probability of committing a type II error is termed β. The **statistical power** of a test (Cohen 1988, Lipsey 1990) is the probability of rejecting a H_0, given that there really is a

genuine effect (i.e. given that H_0 is false). In other words, statistical power = $1 - \beta$. Statistical power generally increases as sample size and effect size increase, and is also dependent on the design of the experiment and the type of test employed (Table 3.1).

Hypothesis testing methods fall into one of two major categories: parametric and nonparametric. **Nonparametric tests** make few assumptions about the underlying statistical distributions and are often used when the errors (residuals) do not conform to the assumptions of a parametric test (**'errors'** are the deviations from the expected values of a statistical model, see below). As a consequence, they are extremely robust to statistical **outliers** (i.e. those data points that are much more extreme than the rest of the measurements in a sample and as a result may cause the sample to seriously violate the underlying assumptions of the statistical model). The main disadvantage of nonparametric tests is that they generally lack the power of equivalent parametric tests (Table 3.1). For example, most nonparametric tests (e.g. Spearman rank correlation, Mann–Whitney *U*-tests, etc.) arrange data into order according to their value and then use their **rank** positions to test for patterns, trends or associations. As a result, information in the data is lost: for example 10, 11, 1000 and 10, 999, 1000 are ranked identically while 11 and 999 have very different values. Similarly, data may be placed in categories and then the frequencies of these categories analysed. Again, information may go unused. The following books examine nonparametric methodology in detail: Meddis (1984), Neave and Worthington (1988), Siegel and Castellan (1988) and Sokal and Rohlf (1995).

Parametric tests assume that data conform to some underlying **error distribution**. Many methods assume that the errors are normally (**Gaussian**) distributed (these are often referred to as ***general* linear models**, as opposed to *generalized* linear models, see below). When the data do not conform to the normal distribution, **transformations** may be applied to raw data to **normalize** the distribution prior to analysis (e.g. the arcsine-squareroot transformation is often used to normalize proportional data).

Many of the classical statistical tests, such as linear regression and analysis of variance, are simply special cases of the general linear model. For example, when the **explanatory terms** (i.e. those terms that explain variation in the data of interest) include a single factor with two levels or categories with equal variances (e.g. treatments A and B), then the test is referred to as a *t*-test; when the factor has more than two levels with equal variance (e.g. three or more treatments), it is referred to as an analysis of variance (**ANOVA**); if there is a single explanatory variable or covariate (e.g. distance from point A), it is referred to as **linear regression**; if there is more than one covariate, it is known as **multiple regression**; if there is a single covariate and one or more factors, it is an analysis of covariance or **ANCOVA**, etc. Thus, it is easier to refer to all of these tests as special cases of a general linear model. These models also allow us to determine whether the responses to explanatory terms are additive or interact in some way. If there is a significant **interaction term** (e.g. A*B), then this indicates that the response to covariate A depends on the level of factor B; in this context, A and B are referred to as **main effects**.

***Generalized* linear models (GLMs)** are generalizations of the linear models referred to as general. They encompass models with normal errors, but may also assume other error distributions (e.g. **Poisson, binomial, negative binomial,**

gamma, etc.). Generalized linear models comprise three components: an **error function**, a **linear predictor** and a **link function**. In sex ratio analyses, we often use GLMs with a *binomial* error function and *logit*-link function. These are often termed **logistic regression models** (see main text).

GLMs, particularly logistic regression models, form the main focus of this chapter. We treat error distributions (i.e. variances) as a consideration during analysis rather than as the focus of the analysis itself; however, sex ratio variances are also of theoretical interest and techniques for their analysis are discussed in Chapter 5.

Some symbols and abbreviations

$-\infty, +\infty$	minus infinity and plus infinity
ln or $\log_e$	natural log (i.e. log to the base e) of x
$\exp(x)$ or e^x	exponent to the power x
χ^2_k	Chi-square (with k degrees of freedom)

3.2.1 Expressing sex ratio data

'Five to one, baby, one in five.
No one here gets out alive'

The Doors

As Jim Morrison's lyric illustrates, *odds* and *proportions* can be used to express the same information (although Morrison stretched poetic licence somewhat since the *odds* 'five to one' are actually equivalent to the *proportion* 'one in six'!). The term 'sex ratio' is commonly used to indicate the numerical relationship between the sexes. However, the quantity of interest is usually expressed as a proportion (conventionally, the number of males divided by the total number of individuals, i.e. males/(males + females)). Here, we conform to this precedent and, unless otherwise stated, we use 'sex ratio' to indicate the proportion of males in a sample, and not a ratio *sensu stricto* (males/females). In Box 3.2 we give an example that shows that analysis of ratios (*sensu stricto*) can lead to errors in interpretation.

3.2.1.1 Proportion data and the binomial distribution

For many organisms, an individual's sex is constrained to be one of two mutually exclusive possibilities: male or female. The data that record this information are said to be *binary*. Other examples include tossing a coin (heads or tails), mortality data (an individual either survives or dies), fertility data (an individual either reproduces or does not) and competition data (an individual either wins or loses). If we examine the sex of one individual and score, for instance, 1 for a male and 0 for a female, the datum is a proportion that has a sample size of one; the sex of the individual is the numerator and the sample size is the denominator, i.e. 1/1 or 0/1 (giving the proportions 1.0 and 0.0).

Often we are interested in determining the average sex ratio or the proportion of males in a group of individuals (e.g. population sex ratio, the sex ratio of the progeny of a given mother, or the sex ratio in a particular brood of offspring). Such data are referred to as *grouped binary data*; the number of males is the numerator, and the total number of individuals sampled is the denominator. For instance, in a brood of six males and seven females the brood sex ratio $= 6/(6 + 7) = 0.46$.

Grouped binary data are often assumed to conform to the *binomial distribution* (Chapter 5) which describes how frequently different sex ratio values are expected. Ungrouped binary data may conform to the *Bernoulli distribution*, a special case of the binomial distribution for sample sizes of one. Our main focus in this chapter is on grouped binary data, as these are most commonly encountered by empiricists, but where differences in the analysis of grouped and ungrouped binary data are apparent, these are highlighted. We explicitly consider ungrouped binary data in sections 3.4.4.3 and 3.5.2.

Box 3.2 Analysis of sex ratios *sensu stricto*: a case history

Leonard and Weatherhead (1996) tested the prediction that parents with high dominance ranks will produce more male-biased offspring sex ratios than low-ranking parents using data from domestic chickens, *Gallus gallus domesticus* (a polygynous bird with stable dominance hierarchies in both males and females). Sex ratios were reported as ratios *sensu stricto* (females/males). Classical ANOVA performed on untransformed data (section 3.3) indicated that sex ratios were not affected by maternal or paternal dominance status. However, a significant effect of mating order on sex ratio was found, using paired *t*-tests (which assume normally distributed error variances, section 3.3), for females that mated with a subordinate male first and later with a dominant male, but not for females that mated with a dominant male first. Leonard and Weatherhead (1996) were unable to propose a simple explanation for this result but concluded that chicken sex ratios are not just a function of random assortment of sex chromosomes and that, given the potential economic value of being able to manipulate chicken sex ratios (a few males are needed for breeding stock but the vast majority are superfluous in agricultural egg production), further exploration would be worthwhile.

We questioned the validity of the analysis since classical ANOVA and *t*-tests were performed on untransformed female/male ratios, and there was no mention of whether error variances were normally distributed. Subsequently, Leonard and Weatherhead (1998) re-analysed these data using sex ratio expressed as proportions (males/(males + females)). Errors were normally distributed and ANOVA and *t*-tests were thus employed without transformation. The previously reported effect of mating order on the progeny sex ratio of females first mated to subordinate males was found to be spurious. Other conclusions were unchanged. The biological conclusion is that there is no consistent bias in chicken sex ratios and that poultry farmers are unlikely to be able to increase productivity by manipulating the status of females' mates. The statistical conclusion is that analyses of ratios *sensu stricto* should be avoided: such ratios are asymmetrical and undefined or infinite if only one sex is present in a clutch, hence mean and variance are not finite (see also Chapter 5) and important information on the size of both the numerator (males) and the denominator (males + females) is lost.

3.2.2 Questions in sex ratio data analysis

Before discussing how to analyse sex ratio data, we briefly consider the questions such analyses are likely to be aimed at addressing. First, it may be of interest to compare observed and expected ratios in order to establish whether an organism has control of its progeny sex ratio. Thus, we may want to ask whether the observed sex ratio differs significantly from the even sex ratio (proportion males = 0.5) that is often taken as the 'null' expectation (e.g. under heterogametic sex determination, Chapter 7). Similarly, it may be of interest to compare an observed distribution of group sex ratios (variances) with the binomial (random) expectation, as this can also indicate sex ratio control and the degree of fit to distributions predicted by evolutionary theory. We briefly summarize methods for testing for sex ratio bias in Box 3.3. Box 3.3 also illustrates a method for analysing sex ratio variance, but this issue is dealt with in detail in Chapter 5.

Second, sex ratio data may be used to explore relationships between sex ratio and specific explanatory terms (factors and covariates; see Box 3.1). Sex ratio theory is a rich and important area of evolutionary biology (e.g. Chapters 1,

Box 3.3 | Comparison of observed and expected sex ratios

Before embarking on a large-scale analysis of sex ratio data, it is often informative to begin by asking two simpler questions: first, does the sex ratio differ from the assumed binomial distribution; and second, does the sex ratio differ from some expected value, such as 0.5. Positive results for one or both of these tests could be indicative of nonrandom variation in the sex ratio distribution. Note that an absence of significant deviation does not necessarily mean that there is no nonrandom variation and that significant deviations do not necessarily indicate parental control of sex allocation, as, for example, sex ratios could be biased due to sexually differential developmental mortality. To illustrate these methods, we use the Example 1 data set (Box 3.5).

Deviation from the binomial distribution

If sex ratio data conform to the binomial distribution, then a GLM with binomial errors (and no explanatory terms other than the intercept, i.e. the null model, Table 3.3) should provide a good fit to the data. We can therefore use the goodness-of-fit test for the null model to determine whether the raw sex ratio data deviate from the binomial distribution (section 3.4.3). To do this, we simply compare the null deviance against the χ^2 distribution with *df* equal to the null degrees of freedom.

Thus, for Example 1, we can ask whether sex ratio distributions for each of the two species (and for both species combined) conform to the binomial distribution

Shirazfish: null deviance = 83.701, null $df = 9$, $P_{(\chi^2_{9=83.701})} < 0.0001$
Merlotfish: null deviance = 32.475, null $df = 9$, $P_{(\chi^2_{9=32.475})} = 0.00016$
Both species: null deviance = 117.961, null $df = 19$, $P_{(\chi^2_{19=117.961})} < 0.0001$

Thus, it appears that both distributions (and the combined distribution) differ significantly from the binomial.

However, when sample sizes are small, this method can severely overestimate the degree of departure from the binomial (Westerdahl *et al.* 1997, Hartley *et al.* 1999). Thus, in these circumstances it is wise to test the robustness of the result by performing *randomization tests*. These involve comparing the deviance of the null model with deviances obtained by a series of *randomly generated* datasets in which 'fish' are allocated to 'samples' at random, while maintaining constant sample sizes. In practice, this requires randomizing fish between samples, while maintaining the same distribution of sample sizes and total number of male and female fish. At each iteration, the deviance of the model is noted and the process is repeated 1000 times. The resulting distribution of deviance values then becomes the null distribution of deviance values against which our model is compared. To determine the significance level of departure from the binomial distribution, we simply divide the number of deviance values greater than or equal to our model's null deviance by 1000 (for *S-Plus* users, a user-defined function for performing these randomizations is available upon request from ken.wilson@stir.ac.uk). However, randomization tests may not perform well when the size of individual samples or the total number of samples is small (Ewen JG, Cassey P & King RAR, unpublished manuscript).

Not surprisingly, given the magnitude of the deviation from the binomial distribution, in this instance, the randomization method confirms the results of our

original analysis and all three distributions were found to be significantly different from the binomial ($P < 0.001$). Figure B3.3a illustrates the distribution of the 1000 randomized null deviances for the Merlotfish dataset. As you can clearly see, the observed null deviance (shown by the solid circle) is significantly higher than any of the values obtained via randomization (histogram). Analysis of sex ratio variances is discussed in detail in Chapter 5.

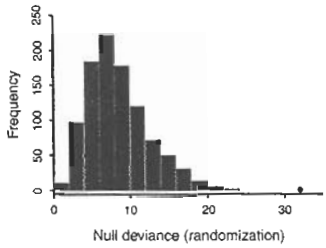

Fig B3.3a Histogram of null deviances obtained from the randomization test. The solid circle represents the observed null deviance.

Deviation from sex ratio equality

In Example 1, there were a total of 638 (253 male and 385 female) Shirazfish sampled, and the overall sex ratio is $253/(253 + 385) = 0.397$. While it is clear that this is not an exact match to 0.5, we need to ask whether the difference is statistically significant. Of the possible tests that can be used, we illustrate five and, as these are widely known and described elsewhere, we do this only briefly.

Binomial test

We could calculate the probability of observing a sex ratio as extreme as 0.397 (i.e. 253 or fewer of one sex in a sample of 638 individuals), assuming that the sex ratio is determined by a random (binomial) process with a mean of 0.5. If this probability is less than 0.05, we conclude that the difference between 0.397 and 0.5 is significant; this is a *binomial test* (see e.g. Siegel & Castellan 1988). As sample size increases, the binomial distribution tends towards the normal distribution, and for samples larger than 35 the normal approximation should be used, but 'corrected' for the fact that the normal distribution is continuous while the binomial distribution involves discrete variables (for details see Siegel & Castellan 1988 p38). Using the normal approximation corrected for discontinuity, an observation as extreme as 253 in a sample of 638 individuals gives $z = -19.99$, $P < 0.0001$; Shirazfish sex ratios are significantly female-biased.

Confidence limits

We could look up the confidence limits for binomial proportions, as published in statistical tables (Rohlf & Sokal 1995), which tell us whether our observed sex ratio falls within the 95% (or 99%) confidence bounds of 0.5 for a given sample size (most statistical packages now offer this facility). If it does, then we can be 95% (or 99%) confident that the observed value does not differ from 0.5 purely due to random sampling error. With a sample size of 638, the lower and upper 99% confidence limits for 0.5 are 0.45 and 0.55 respectively. As our observation of 0.39 is outside these bounds, we can be confident that the bias in the Shirazfish population sex ratio is significantly greater than expected by chance alone.

Simulation

Another way to address this same problem is to determine the relative confidence intervals by simulation. In other words, we generate a large number (i.e. > 1000) of simulated datasets comprised of random samples drawn from the binomial distribution with the mean equal to 0.5 and sample size equal to the number of individuals in our dataset (638 in our particular example). If our observed sex ratio lies outside the appropriate confidence intervals for our simulated dataset, then the sex ratio is significantly different from 0.5. This approach is illustrated in Figure B3.3a, in which the histograms represent the simulated dataset and

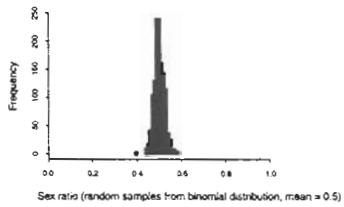

Fig B3.3b Histogram of simulated sex ratios generated by randomly sampling from a binomial distribution with mean equal to 0.5 and sample size equal to 638. The solid circle indicates the observed sex ratio.

the solid circle represents the observed sex ratio. As you can see, none of the 1000 simulated datasets had a sex ratio that was as low as that which we observe, indicating that the probability of observing this sex ratio by chance alone is less than 1 in 1000, i.e. $P < 0.001$.

Chi-square goodness-of-fit test

We could compare the observed numbers of males (253) and females (385) with the number of each sex expected under sex ratio equality (638/2 = 319 individuals of each sex) using a chi-square test, which is based on the *deviations* of observed from expected values (for details see e.g. Siegel & Castellan 1988 p45, Sokal & Rohlf 1995 p695). The chi-square, X^2, value computed is also known as *Pearson's statistic* to distinguish it from the chi-square sampling distribution, χ^2, which it approximates. The Shirazfish data generate a Pearson's statistic of 27.31 with $df = 1$, which is greater than the critical value in χ^2 tables for $P = 0.001$, so we conclude that the sex ratio is significantly female-biased. With small samples, and with biased sex ratio expectations, the expected value of one or both sexes may be five or less. In such cases *Fisher's exact test* should be used instead of the chi-square test (e.g. Siegel & Castellan 1988 p103, Crawley 1993 p237).

Likelihood ratio goodness-of-fit test

We could compare the observed and expected numbers of males and females with the numbers expected under sex ratio equality using a *G*-test which is based on the *ratios* of observed and expected values (e.g. Crawley 1993 p234, Sokal & Rohlf 1995 p688, Zar 1999 p505). The Shirazfish data generate $G = 26.87$ with $df = 1$, which is greater than the critical value in χ^2 tables for $P = 0.001$, so again we conclude that the sex ratio is significantly male-biased. Note also that G and X^2 values are generally similar.

Choice of test

Which of these five tests should be preferred is determined by the power function for the class of alternative hypotheses under consideration (E. Meelis pers. comm.). However, the binomial test will usually be the definitive test, the *G*-test is generally preferred over the chi-square goodness-of-fit test (Crawley 1993, Sokal & Rohlf 1995, but see Zar 1999) and the confidence interval and simulation will generally give similar results for large sample sizes.

13, 19 & 20) and there are many predictions that can be tested in this way. Analysis of such relationships forms the main focus of this chapter.

3.3 Classical analyses of proportion data

A variety of different methods has been used to analyse sex ratio data in recent years (Box 3.4). In this section we review some of the more traditional methods, highlighting their strengths and weaknesses. We illustrate these points using a fictitious dataset (Box 3.5) on the effects of a pollutant on the sex ratios of two fish species in an Australian creek. The first thing we need to do is to plot the data (Figure 3.1a). The figure appears to indicate that, in both species, populations close to the source of pollution tend to have female-biased sex ratios and that as we get further away from the pollution source the sex ratio

Box 3.4 Survey of statistical approaches used in recent sex ratio literature

Studies on sex ratio are published in a range of journals within the general field of evolutionary ecology, as well as in taxon-specific journals. To assess which statistical methods are the most commonly used for analyses of sex ratios and other proportional data, we surveyed empirical studies on sex ratio and closely related issues (e.g. sex determination, sex allocation, sex-biased mortality) published in 1994–2000 in four leading evolutionary, ecological and behavioural journals. We found 83 studies, some of which employed more than one approach: see Table B4.3. Part (a) of the table scores the methods used to test for departure from some expected sex ratio value (e.g. 0.5) (see Box 3.3) and part (b) scores methods used to examine trends in sex ratio with explanatory variables, which is the main focus of this chapter.

Table B3.4 Recently used methods

	Journal (Number of studies)				
Sex ratio analysis method	Animal Behaviour (34)	Behavioral Ecology (16)	Evolution (18)	Oikos (15)	Totals (83)
(a) Deviation from expected sex ratio					
No statistical test			2		2
Binomial test					
Fisher's exact test	2	2	4	3	11
χ^2-test	10	2	4	6	22
G-test	5	1	3	2	11
Other	1		2	3	6
(b) Relationships with explanatory variable(s)					
1. No statistical test	3		1	1	5
2. Nonparametric tests	11	6	3		20
Standard parametric tests:					
3. No transformation	5	3	2	1	11
4. Arcsine squareroot transformation	11	3	7	4	25
5. Other transformation	2	1	1		4
6. Generalized linear modelling (logistic)	4	5	3	4	16

Note that some authors who used standard parametric tests without transformation first tested the appropriateness of the assumption of normal error variances, while others attempted to use GLMs but found a degree of overdispersion to be too large (e.g. heterogeneity factor >4) and opted to use standard parametrics following arcsine squareroot transformation instead (e.g. Flanagan *et al.* 1998). However, some authors employed standard techniques despite using statistical

packages (e.g. *SAS*) which are capable of running GLMs; possibly because they were unaware of the advantages of GLM analysis?

We explored our survey data by ranking methods in part (b) in rough order of 'advancement' from 1 (no statistical analysis) to 6 (logistic GLMs) and gave each study a 'sophistication score' equal to the rank of the most advanced method used. We found no evidence that the level of 'sophistication' changed significantly during the surveyed years, or that it is related to the journal in which the study was published (Year, $\chi^2_1 = 0.34$, $P > 0.1$; Journal, $\chi^2_3 = 0.46$, $P > 0.1$; results from log-linear analysis, which is appropriate for count data, Crawley 1993). We also found no relationship with year or journal when we carried out (binary) logistic analyses on: (1) the proportion of parametric tests out of all methods in part (b) (mean = 0.799, Year, $\chi^2_1 = 0.23$, $P > 0.1$; Journal, $\chi^2_3 = 2.77$, $P > 0.1$), (2) the proportion of logistic GLMs out of all methods (mean = 0.277, Year, $\chi^2_1 = 2.37$, $P > 0.1$; Journal, $\chi^2_3 = 4.72$, $P > 0.1$), and (3) the proportion of logistic GLMs out of all parametric methods (mean = 0.346, Year, $\chi^2_1 = 2.53$, $P > 0.1$; Journal, $\chi^2_3 = 3.25$, $P > 0.1$). We conclude that GLMs are underused and that the situation has not recently been improving.

Box 3.5 Example 1: Pollution and sex ratios in Australian fish

Our first data set is a hypothetical example, in which we examine the effects of a pollutant (alcohol) on sex ratios in two imaginary fish species in Stubbie Creek, Australia: the wide-mouthed Shirazfish and the big-nosed Merlotfish. We imagine that the data were collected by netting fish at 100-m intervals along the creek for a distance of up to 1000 m from the source of the pollutant. The hypothesis we are testing is that the pollutant leads to biased sex ratios in both species.

Table B3.5

Distance from pollution source (m)	Shirazfish		Merlotfish	
	sample size (no. fish)	sex ratio (proportion males)	sample size (no. fish)	sex ratio (proportion males)
100	67	0.30	7	0.14
200	120	0.50	12	0.33
300	21	0.33	30	0.40
400	103	0.13	5	0.20
500	88	0.65	46	0.59
600	34	0.29	67	0.25
700	99	0.31	29	0.31
800	34	0.38	74	0.43
900	67	0.57	35	0.71
1000	5	0.80	134	0.48

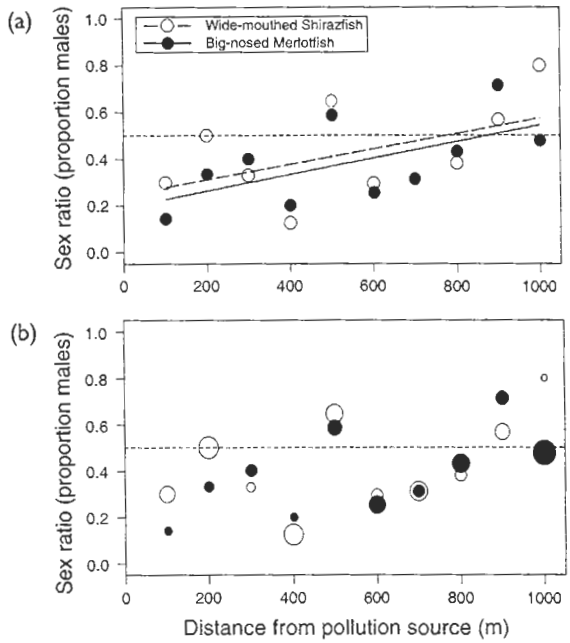

Fig 3.1 Relationship between sex ratio and distance from pollution source for two fish species in Stubbie Creek (Example 1, Box 3.5). The solid line is the least-squares fit to the Merlotfish dataset and the dashed line is the fit to the Shirazfish. (b) Symbol size is proportional to the sample size upon which the sex ratio is based. On both panels the dotted line shows sex ratio equality (0.5).

gets closer to 0.5. We will start with nonparametric analyses (section 3.3.1.2) and then go on to linear models with normal errors (sections 3.3.2 & 3.3.3). In sections 3.4 and 3.5, we analyse these data using generalized linear models.

3.3.1 Nonparametric tests

Nonparametric tests (e.g. Mann–Whitney *U*-test, Kruskal–Wallis test, Spearman's correlation) are frequently employed in the behavioural sciences because they are simple to implement by hand or by computer and because they make no assumptions about the shape of the underlying error distribution and thus they are extremely robust to outliers (Box 3.1). This does not mean that these tests are 'assumption-free', however, since most nonparametric tests usually assume that the observations are independent and sometimes that the variable under study has underlying continuity or that the distributions have similar shape across groups. Nevertheless, the assumptions associated with nonparametric tests are fewer and weaker than those associated with equivalent parametric tests (Box 3.1). As a consequence, if all of the assumptions of a parametric test are met, nonparametric tests lack *power* (Box 3.1) and are *wasteful* (Siegel & Castellan 1988).

3.3.1.1 Power-efficiency

The degree of wastefulness of a test can be expressed by its *power-efficiency*, which is concerned with the increase in sample size required to make test B (e.g. a nonparametric test) as powerful as test A (e.g. an equivalent parametric test), when the significance level is held constant and the sample size of test A is held constant. Thus

$$\text{Power-efficiency of test B (\%)} = 100 \times N_A/N_B. \qquad \text{(eq. 3.1)}$$

N_A and N_B are the relative sample sizes required to give test B the same power as test A. For example, if test B requires a sample size of $N_B = 25$ to have the same power that test A has when it has a sample size of $N_A = 20$, then test B has a power-efficiency of $100 \times (20/25) = 80\%$ (Siegel & Castellan 1988). In other words, test A would be just as effective with a sample that was 20% smaller than that used in test B. Table 3.1 compares the power-efficiencies of some of the commonly employed nonparametric tests with their most comparable parametric test.

3.3.1.2 Example 1: Fish sex ratios

Now let's return to Example 1 (Box 3.5). We want to investigate whether sex ratio varies consistently with distance from the pollution source in the two fish species. There are a number of ways that we can employ nonparametric tests. If we use Spearman rank-order correlations to assess whether there is an association between sex ratio and distance from the pollution source for each fish species the answer appears to be 'no' (Shirazfish: $r_s = 0.467$, $n = 10$, $P = 0.167$; Merlotfish: $r_s = 0.624$, $n = 10$, $P = 0.063$). Note that if we were testing an *a priori* hypothesis, for example based on data showing that males were more susceptible to the pollutant than females, we could argue that one-tailed probabilities were

Table 3.1 Power-efficiency of some of the commonly employed nonparametric tests

Nonparametric (*NP*) test	Parametric (*PP*) test	Power-efficiency (%) of *NP* test	Comments
Spearman's rank-order correlation	Pearson's product-moment correlation	91%	Power-efficiency same as for Kendall's rank-order correlation
Wilcoxon signed ranks test	paired *t*-test	<95.5%	Power-efficiency ~95% even for small sample sizes
Wilcoxon Mann–Whitney *U*-test	*t*-test	<95.5%	Power-efficiency ~95% for small sample sizes
Kolmogorov–Smirnov two-sample test	*t*-test	<95%	Power-efficiency declines slightly with increasing sample size
Kruskal–Wallis one-way ANOVA	One-way ANOVA (*F*-test)	<95.5%	
Friedman two-way ANOVA	Two-way ANOVA (*F*-test)	64% ($k = 2$) to 91% ($k \geq 20$)	Power-efficiency dependent on number of matched samples (k)

more appropriate (Box 3.1) and the significance levels would be reduced to $P = 0.083$ and $P = 0.031$, respectively (see also section 3.5.4). Notice that, in both cases, the correlation coefficients are fairly large ($r_s \geq 0.46$), and it seems likely that the lack of significance for these two relationships is due to low statistical power (see Box 3.1 and below). The power-efficiency of the Spearman's rank correlation test is 91% when compared to the most powerful parametric correlation test (Pearson's product-moment correlation; Table 3.1). Re-analysing Example 1 data using Pearson's correlation following arcsine-squareroot transformation to help normalize the error distribution (Section 3.3.3.2, Box 3.6) yields the following correlation coefficients: Shirazfish: $r = 0.498$, $df = 8$, two-tailed $P = 0.143$, Merlotfish: $r = 0.618$, $df = 8$, two-tailed $P = 0.057$. Thus, there does not appear to be a significant relationship between sex ratio and distance from a known pollution source, regardless of whether we use a parametric or non-parametric correlation test. But, was the power of our analysis (Box 3.1) great enough to be able to detect a significant relationship if there was one? Ideally, the power of our test should be greater than 80%. The statistical power of a test is determined by sample size, the amount of variation in the data and the magnitude of the effect one is trying to detect. We can determine the power of our two correlations using the following formula (Cohen 1988, Zar 1999)

$$Z_\beta = (z - z_\alpha)\sqrt{(n-3)},$$

where z and z_α are the Fisher transformations for r (the correlation coefficient) and r_α (the critical value of r), the Fisher transformation $= 0.5 \ln(1 + r/1 - r)$, n = sample size, and Z_β is the probability of the normal deviate, which can be translated into power ($1-\beta$) by comparing against the appropriate tabulated value (e.g. Appendix Table B.2 in Zar 1999).

These days, a simper way to determine power is to use a power calculator (such as that which can be found at http://ebook.stat.ucla.edu/calculators/powercalc/). Using this calculator, the power of our two Pearson's correlation tests were determined to be 30.4% and 48.0%, respectively, which are nowhere near the desired 80%. These calculators can also be used to determine the sample sizes required to achieve a given power. In this example, sample sizes of 30 and 19, respectively, are required to achieve 80% power.

Box 3.6 | Effect of arcsine-squareroot transformation

Here we illustrate the effect of arcsine transformation on proportional data. Figure B3.6a shows the relationship between proportions and their transformed values. Arcsine transformation has the effect of stretching out the ends of the distribution, such that the truncation that occurs when the mean of the (binomial) distribution is close to zero or one is reduced. As a consequence, the distribution should become more normalized under transformation.

Figure B3.6b shows the distribution of binomially distributed data, for a range of clutch sizes (CS = 2, 5, 10 and 20) and mean sex ratios (mean SR = 0.5, 0.75 and 0.9). The data show the frequency distribution of 5000 random samples taken from the binomial distribution, using the *rbinom* function in *S-Plus* (open bars) and the effect of arcsine-squareroot transformation on the distribution (closed bars). Note that the effects of arcsine transformation on mean SR = 0.25 and 0.1 are equivalent to those illustrated for mean SR = 0.75 and 0.9, respectively.

Although the nontransformed sex ratios are approximately normal for sex ratios close to 0.5 (especially when clutch sizes are large), the data are severely skewed when sex ratios are heavily biased towards one or other sex (especially when clutch sizes are small). Arcsine transformation tends to make the data more normal (cf. the open and closed bars in the bottom-right figure), though in some cases the effect is to make the data less normal (cf. middle-right distributions). For small clutch sizes and heavily biased sex ratios, arcsine transformation fails to normalize the data.

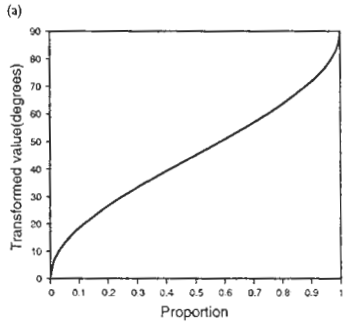

Fig B3.6a Effect of arcsine transformation on proportional data.

Lack of power is not the only problem with the analysis described above. Another is that it fails to take into account the fact that we appear to have the same relationship in both datasets. Ideally, we would want to perform a test in which we utilize information from both species simultaneously and ask whether we get the same relationship in both. Unfortunately, there is no easily accessible nonparametric test that is equivalent to analysis of covariance (but see the Page Test for Ordered Alternatives, Siegel & Castellan 1988). We could perform a Wilcoxon signed ranks test (equivalent to a paired *t*-test) to determine whether the sex ratio variation *within* our two fish species is greater than that *between* them, but this would tell us nothing about their respective sex ratio trends along the creek. An alternative procedure is to perform a Fisher combined probability test (Fisher 1954, section 21.1; Box 18.1 Sokal & Rohlf 1995) that allows us to use the probabilities derived from the correlations we carried out on the two species.

The calculation of the Fisher's combined probability estimate is based on the fact that $-2\ln P$ is distributed as χ^2_2 (see Box 18.1 in Sokal & Rohlf 1995). Thus, by evaluating twice the negative natural logarithm of each of the (k) probabilities we wish to combine, and summing them, we obtain a total ($-2\Sigma\ \ln P$) that can be compared against the χ^2 distribution with $2k$ (= 4, in this example) degrees of freedom (i.e. χ^2_{2k}). In our example, based on the one-tailed P-values from the Spearman rank correlations

$$\begin{aligned} -2\Sigma\ \ln P &= -2 \times (\ln 0.083 + \ln 0.031) \\ &= -2 \times (-2.4889 - 3.4737) \\ &= -2 \times -5.9626 \\ &= 11.925. \end{aligned} \qquad \text{(eq. 3.2)}$$

When compared with χ^2_4, this yields a two-tailed probability of $P_{\text{combined}} = 0.0358$. Thus, when we use the information we have on the two fish species, there appears to be a significant trend for sex ratio to increase with distance from the pollution source, but the statistical evidence for such a relationship is far from convincing. As indicated above, this is probably due to a lack of statistical power, because (1) we are relying

(b)

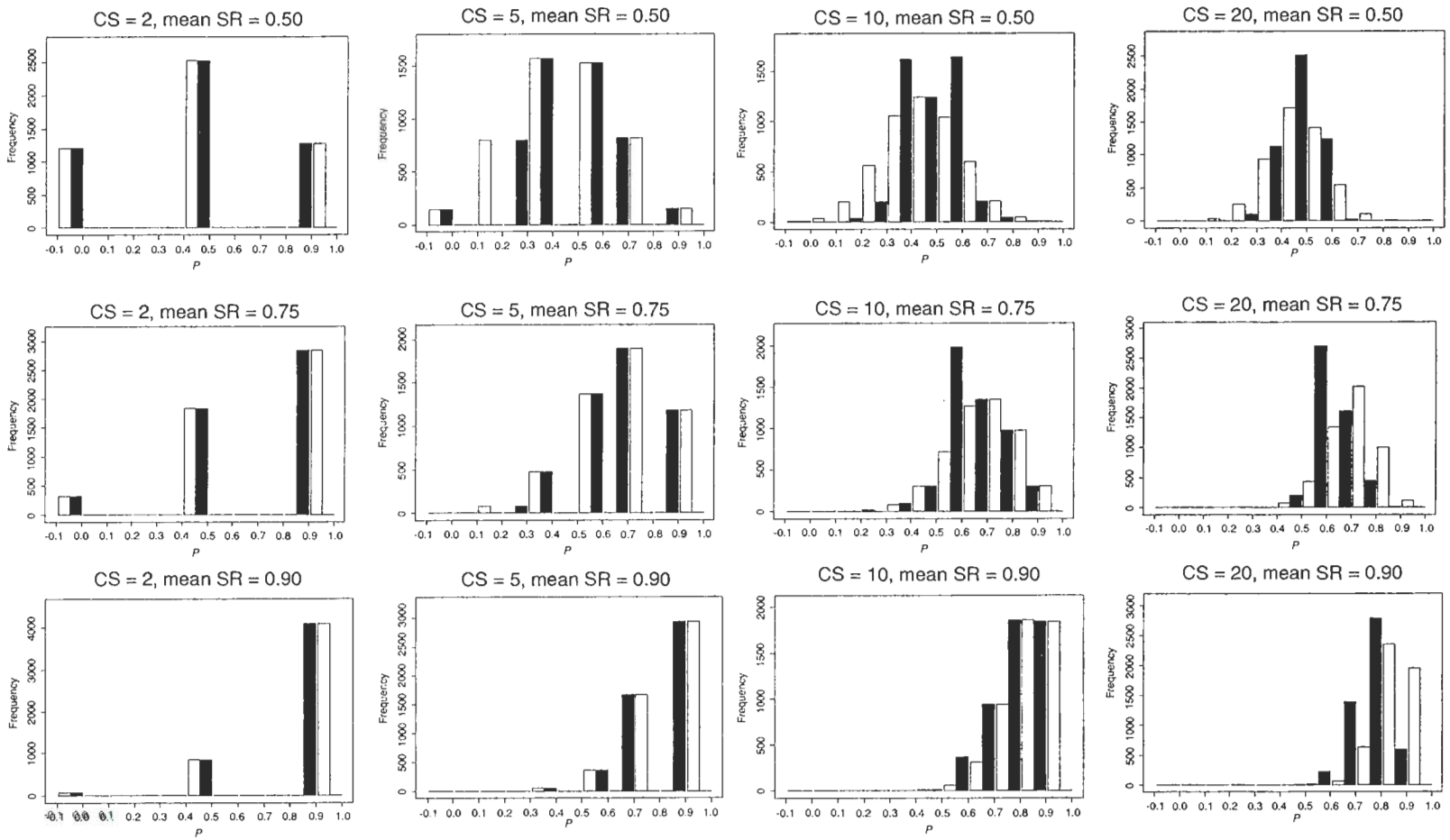

Fig B3.6b Effect of arcsine transformation on randomly generated binomial datasets. Open bars indicate the observed binomial data generated by randomly sampling from a binomial distribution with sample size (i.e. clutch size = CS) equal to 2, 5, 10 or 20, and sample mean (i.e. mean sex ratio = mean SR) equal to 0.5, 0.75 or 0.9. The closed bars are the arcsine-transformed proportions.

on nonparametric tests that do not allow us to adequately combine factors and covariates in the same model, and (2) we are losing information about sample sizes. In the following sections, we address each of these concerns. The first deficiency is covered by considering (parametric) linear models (section 3.3.2), and the second by considering weightings within these models (section 3.3.3.4). We will then move on to consider more carefully the underlying assumptions of these models (section 3.4).

3.3.2 General linear models

General linear models are parametric models which assume that the underlying error distribution is normal (Gaussian). They are a special type of generalized linear model, which is discussed fully in section 3.4. They include most of the 'classical' methods that most readers will be familiar with, including linear regression, analysis of variance (ANOVA) and analysis of covariance (ANCOVA).

Recall that the question we are trying to address is: does sex ratio vary consistently with distance from the pollution source in our two species of fish? We could perform separate linear regressions for the Shirazfish and the Merlotfish, but it makes better sense to use all of the data and perform an ANCOVA in which, effectively, we are asking: does the relationship between distance from the pollution source and sex ratio differ between our two species. An ANCOVA on these data (first, third and fifth columns of Table B3.5) generates the following ANOVA table (the output comes from *S-Plus*):

```
Terms added sequentially (first to last)
                 Df  Sum of Sq  Mean Sq  F Value    Pr(F)
Distance          1    0.1931   0.1931   6.9151   0.0182
Species           1    0.0083   0.0083   0.2974   0.5930
Distance:Species  1    0.0002   0.0002   0.0081   0.9293
Residuals        16    0.4470   0.0279
```

Thus, it appears that sex ratio does not vary between species ($P = 0.59$) and that the relationship between distance from the pollution source and sex ratio does not vary between the two species ($P = 0.92$), but that sex ratio does vary (increase) with distance ($P = 0.018$). The first line of this output reminds us, however, that these results are based on *sequential* sums of squares, so the order in which the explanatory variables appear in the model may influence the results. We therefore need to undertake model simplification (section 3.4.5). After simplification, it appears that our 'best' (most parsimonious) model is one in which `Distance` is the only significant explanatory term:

```
Terms added sequentially (first to last)
           Df  Sum of Sq  Mean Sq  F Value    Pr(F)
Distance    1    0.1931   0.1931   7.6337   0.0128 **
Residuals  18    0.4555   0.0253
```

And the relationship is described by the following regression line

$$\texttt{Sex ratio} = 0.2173 + 0.0003 \times \texttt{Distance}. \qquad \text{(eq. 3.3)}$$

There are a number of problems with this analysis, however. First, since our data are proportions, the assumption of normal errors made by classical regression methods is likely to be violated, particularly when proportions are less than 0.3 and greater than 0.7 (Zar 1999). Indeed, inspection of the *normality plot* (see section 3.4.6) for this linear regression shows that there is considerable curvature in the residuals, indicating significant deviation from normality (check it yourself!). A common 'quick-fix' for this problem is often to perform some sort of *transformation*.

3.3.3 General linear models with transformed data

3.3.3.1 Probit transformation

One of the earliest transformations applied to binomial proportion data was the *probit transformation*, which has most commonly been employed in the analysis of dose–response data from bioassays. Probit transformation evolved when such analyses were performed by hand using probit paper. With the advent of desktop computers, this method is now considered rather old-fashioned. For further information see Finney (1971) and Crawley (1993).

3.3.3.2 Arcsine transformation

A much more commonly employed transformation used by sex ratio biologists is known as the *arcsine-squareroot transformation* (also known as the *arcsin transformation* or *angular transformation*). This involves taking the square-root of the proportion, p, and transforming it to its arcsine (i.e. the angle whose sine is $\sqrt{p}$)

$$p' = \arcsin \sqrt{p} \qquad \text{(eq. 3.4)}$$

For proportions between 0 and 1, the transformed values will range between 0 and 90 degrees (some statistical tables and packages present the transformation in terms of radians; a radian is $180^\circ/\pi = 57.2958$ degrees). Note that prior to arcsine transformation, the data must be represented as proportions and not as percentages. Arcsine transforming the fish sex ratio data (Example 1) has little impact on the results of our analysis

```
Terms added sequentially (first to last)
           Df  Sum of Sq  Mean Sq  F Value   Pr(F)
 Distance   1    0.2239   0.2239   7.7734  0.0121 **
Residuals  18    0.5185   0.0288
```

And the relationship (in degrees) is described by the following regression line

$$\texttt{Sex ratio} = 27.6331 + 0.0211 \times \texttt{Distance}. \qquad \text{(eq. 3.5)}$$

Whilst the arcsine transformation often helps to normalize proportion data, it does not work well at the extreme ends of the distribution, i.e. near 0 and 1 (Box 3.6); of course, this can be checked by producing a normality plot (see section 3.4.6). Moreover, arcsine transformation does not get around another major attribute of proportion data, namely that the responses are strictly bounded between 0 and 1 (or 0% and 100%). Thus, the classical linear methods that we used earlier (i.e. linear regression and ANCOVA) could easily predict biologically unrealistic or even impossible results, especially if the variance is high and the data lie close to zero. In Example 1, the linear regression line describing the relationship between sex ratio and distance from the pollutant source indicates that at a distance of 2609 m (untransformed data) or 2956 m (arcsine-transformed data), the creek will comprise only males, and at greater distances the sex ratio will exceed 1! Although extrapolating so far beyond the observed data would be ludicrous, the point remains that classical linear models can predict values that lie outside biologically sensible bounds.

3.3.3.3 Logistic transformation

One way round this problem is to apply the *logistic transformation*, in which our *success probability* p (i.e. proportion of males in our sample) undergoes the following transformation, written as logit (p)

$$\text{logit}\,(p) = \ln\left(\frac{p}{1-p}\right). \qquad \text{(eq. 3.6)}$$

Thus, for p in the range 0 to 1, logit (p) will range between $-\infty$ and $+\infty$, respectively. If we apply the logit transformation to a simple linear model, we produce the following linear logistic model

$$\text{logit}\,(p) = \ln\left(\frac{p}{1-p}\right) = a + bx. \qquad \text{(eq. 3.7)}$$

Note that $p/(1-p)$ is the statistical *odds* of success (Jim Morrison's 'five to one'), and so the logistic

transformation of p is the *log odds* of success. We can make use of this fact to re-write eq. 3.7 to make p a function of x

$$p = \frac{e^{(a+bx)}}{1 + e^{(a+bx)}}. \qquad \text{(eq. 3.8)}$$

Thus, when $x = -\infty$, $p = 0$ and when $x = +\infty$, $p = 1$, so fulfilling our need for p to be strictly bounded between 0 and 1.

If we apply the logit transformation (eq. 3.7) to the Example 1 sex ratio data and perform an ANCOVA, we arrive at the following result

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
Distance	1	4.2161	4.2161	7.8672	0.0117 **
Residuals	18	9.6465	0.5359		

And the relationship (*in logits*) is described by the following regression line

$$\texttt{Sex ratio} = -1.3102 + 0.0016 \times \texttt{Distance}. \qquad \text{(eq. 3.9)}$$

Back-transforming eq. 3.9 (using eq. 3.8), the predicted sex ratio at the pollution source is $e^{-1.3102}/(1 + e^{-1.3102}) = 0.2124$ and even at 10 000 m away from the pollution source, the predicted sex ratio remains within realistic bounds (e.g. $e^{(-1.3102+0.0016\times10\,000)}/[1 + e^{(-1.3102+0.0016\times10\,000)}] = 0.9999996$).

As we shall see in section 3.5, the logit transformation forms the basis for *logistic regression* (i.e. generalized linear modelling with binomial errors and logit link function). We do this within the GLM context (rather than using simple linear regression, as above), because: (1) logistic regression allows for the nonconstant binomial variance (the variance of the binomial distribution equals $np(1 - p)$ and peaks at $p = 0.5$); (2) it deals with the fact that logit(p) values near 0 or 1 are infinite; and (3) it allows for differences between sample sizes by weighting the regression (Crawley 1993, 2002).

3.3.3.4 Weighted linear regression

In all of the models we have considered so far, each data point (i.e. sex ratio) contributes equally. However, it is clear that if we have two sex ratios, and one is based on a sample of five individuals and the other on 500, we should have much more confidence in the value derived from the larger sample. Thus, in those cases where it is known *a priori* that not all observations contribute equally to the fit of the model, we should *weight* our observations according to the confidence we have in them (usually some function of sample size). This process is called *weighted regression*.

Figure 3.1b shows the data for Example 1 with the size of the symbols reflecting the size of the sample upon which the sex ratio was estimated (i.e. the denominator of the sex ratio; the second and fourth columns in the Table B3.5). It is very clear that there is considerable sample size variation between the data points. A weighted ANCOVA on the logit-transformed data finds that neither `Species` nor `Distance` (nor their interaction) is statistically significant (e.g. for `Distance`, $P > 0.16$). Thus, when we weight sex ratios according to sample size, it appears that there is no consistent relationship between sex ratio and distance from the source of pollution. This is because most of the extreme sex ratios (i.e. those that deviate most from 0.5) are based on small sample sizes.

In section 3.4 we incorporate the ideas of weighted regression and logit transformation into a technique known as generalized linear modelling.

3.4 | Generalized linear models

The *general* linear models we discussed in the previous section are based on the underlying assumption that the distribution of residuals around the fitted model (i.e. the error distribution) is Gaussian (= 'normal'), and that these residuals show no systematic variation with respect to the mean (i.e. that the variance is constant). However, these two assumptions are often violated (as we have seen already for sex ratio data). *Generalized linear models* (GLMs) differ from

the *general* linear models encountered previously in allowing one to also specify non-normal error variances, such as Poisson, binomial, negative binomial, gamma and exponential (section 3.4.2.1). We can use GLMs to analyse sex ratio data, but before we can do that we need to understand the rationale and some of the benefits of the GLM approach.

3.4.1 The pros and cons of using GLMs

Generalized linear modelling provides a single theoretical framework for analysing many different types of data. This makes it an extremely powerful and flexible approach. Also, by careful choice of an appropriate *link function* (section 3.4.2.3), the GLM will constrain the predicted values to lie within realistic bounds (as with the logistic transformation, section 3.3.3.3). The main limitation of the GLM approach is that it is restricted to models that are *linear*. This does not mean that GLMs can be used only to describe straight-line relationships, but that it must be possible for the model to be structured in such a way that it describes a linear relationship. For example, the following nonlinear equation

$$y = e^{(a+bx)} \qquad \text{(eq. 3.10)}$$

can be linearized by log-transforming both sides

$$\ln(y) = a + bx. \qquad \text{(eq. 3.11)}$$

Within GLMs, this process is performed by $\log_e$ transforming the dependent variable by specifying a log *link function* (section 3.4.2.3). Some models are intrinsically nonlinear because there is no transformation that can linearize them in all parameters. For example

$$y = a + \frac{b}{c + x}. \qquad \text{(eq. 3.12)}$$

In these circumstances, the GLM is unable to estimate all of the parameters (a, b and c) and we must undertake nonlinear modelling.

3.4.2 Components of a GLM

Generalized linear models have three essential ingredients (Crawley 1993, 2002 provides a fuller explanation).

3.4.2.1 Error structure

The error structure describes the shape of the distribution of residual values around the fitted model. Classical linear models assume a normal (Gaussian) distribution; GLMs allow other error distributions to be defined such as Poisson errors (e.g. for count data), negative binomial errors (e.g. for parasite load data), exponential errors (e.g. survival times) and binomial errors (e.g. sex ratios, mortality and other proportion data).

3.4.2.2 Linear predictor

The linear predictor is a linear equation defining the relationship between the predicted y values and one or more explanatory variables, *on the scale determined by the link transformation* (section 3.4.2.3). The number of terms in the linear predictor is the same as the number of parameters to be estimated from the data. So, for a simple linear regression, there are two terms in the linear predictor (slope and intercept). To determine the fit of a given model, the GLM evaluates the linear predictor for each value of the response variable and compares this with a *transformed* value of y that is determined by the link function. The fitted value is determined by *back-transforming* the predicted values to the original scale (so, for example, with a *log link*, the fitted value is the *antilog* of the linear predictor and with the *reciprocal* link it is the *reciprocal* of the linear predictor). This will become clearer when we go on to examine a specific example (section 3.5.1.1).

3.4.2.3 Link function

Data on proportions, such as sex ratios, are frequently described by the *logistic curve* (Figure 3.2a), because this equation (eq. 3.13) asymptotes at 0 and 1 (or 0% and 100%), and guards against unrealistic values being predicted.

$$p = \frac{e^{(a+bx)}}{1 + e^{(a+bx)}}. \qquad \text{(eq. 3.13)}$$

Clearly, eq. 3.13 describes a nonlinear relationship. However, it can be linearized by applying the *logit transformation* we encountered earlier

$$\ln\left(\frac{p}{1-p}\right) = a + bx. \qquad \text{(eq. 3.14)}$$

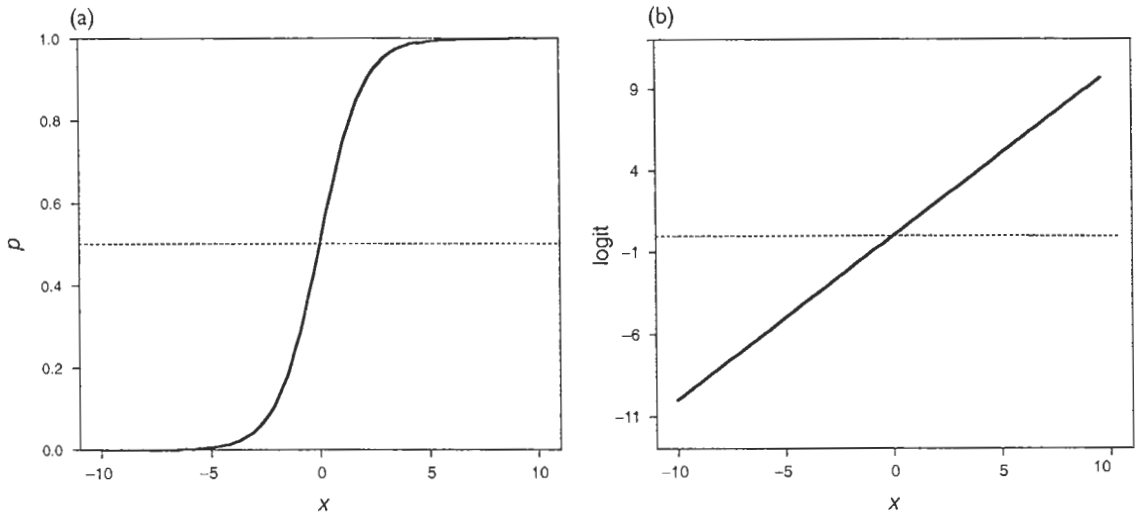

Fig 3.2 The logistic curve (a) can be used to ensure that the fitted values lie between 0 and 1, and the logit transformation (b) can be used to linearize the relationship.

This transformation is known as the *link function*, and it relates the mean value of y to its linear predictor (section 3.4.2.2). It is essentially just a transformation that linearizes the model and ensures that the fitted values stay within reasonable bounds (Figure 3.2). As indicated in section 3.4.2.2, the values that emerge from the linear predictor are on the scale of the link function, and predicted values of y are generated by back-transforming the linear predictor to the original scale. So, for example, to ensure that predicted count data (e.g. number of beetles per quadrat) never become negative, a *log link* function would be applied. This is because the fitted values would then be antilogs of the linear predictor, and all antilogs are greater than or equal to zero. In the case of proportion data, the *logit link* function is generally applied to ensure that predicted values never exceed one or drop below zero (other link functions include the *identity link* for normal errors, the *reciprocal link* for gamma errors, the *probit link* for bioassays and the *complementary log-log link* for dilution assays). Although the default link function for binary and binomial data is the logit link, the (asymmetrical) complementary log-log link should also be assessed as it will sometimes lead to a lower residual deviance (Crawley 1993, for an empirical example see Petersen & Hardy 1996).

3.4.3 Determining the best-fit model: maximum-likelihood

The 'classical' methods with which most of us are familiar (linear regression, ANOVA, etc.) utilize *least-squares* (LS) methods for determining the best-fit model. In other words, we find the model that *minimizes the sum of squares* of the departures from the observed y values from their predicted values. In contrast, generalized linear modelling determines the best-fit model by *maximum-likelihood* (ML) methods. When the GLM has normal errors and an identity link function (as in 'classical' models), ML and LS give identical results (indeed, linear LS methods are a subset of ML, in much the same way as *general* linear models are a subset of generalized linear models). For other kinds of error structure and link functions, LS methods may produce biased parameter estimates, and so ML is generally preferred (e.g. McCullagh & Nelder 1989). Even though ML estimation is relatively straightforward, it is rather laborious, and so most biologists are happy to treat the process as a 'black box'. Those interested in the mechanistic basis to ML estimation in a biological context can find examples in Crawley (1993, 2002) and McCallum (2000).

The basic idea behind any statistical modelling procedure is to determine the parameter values that lead to the best fit of the model to the data. With LS regression, the best-fit model is determined by minimizing the residual sum of squares. With ML, we ask: given our data and our choice of model, *what parameter values*

Table 3.2 A description of deviance terms

Deviance term	Description
Null deviance	Deviance associated with the *null model* (≡ total sum of squares)
Residual deviance (deviance)	Deviance remaining after some or all terms have been included in the model (≡ residual sum of squares)
Change in deviance	Deviance associated with inclusion of a particular term in a model (≡ sum of squares for a particular term)

Table 3.3 Terminology for stages of model simplification

Model	Description
Saturated (full) model	Perfect fit; zero deviance and *df*; one parameter for each observation
Maximal model	Contains all factors, interaction terms and covariates under consideration
Current model	The current model; number of parameters ≤ maximal and ≥ minimal model
Minimal model	A model with minimal number of terms, in which all parameters are significantly different from zero, and no important terms have been omitted
Null model	Only the grand mean (i.e. one parameter) is fitted; deviance ≡ total sum of squares in 'normal' models

maximize the likelihood of the data being observed (hence the term 'maximum likelihood')? *Likelihood* (or more commonly *log-likelihood*) is used here in a formal sense for assessing the statistical *odds* of producing a particular outcome. The 'best' model is therefore the model that produces the minimal *residual deviance* (Table 3.2), subject to the constraint that all the parameters in the model should be statistically significant (Table 3.3).

Residual deviance is twice the difference between the maximum achievable log-likelihood (i.e. that obtained when the predicted and observed values are identical) and that attained by the model under consideration (McCullagh & Nelder 1989). For most error structures, deviance is distributed asymptotically as *chi-square* (χ^2) and so the *goodness-of-fit* of a model can be determined by calculating the deviance and testing it against the chi-square distribution with the appropriate degrees of freedom (*df*). By convention, if $P > 0.05$, then we usually declare that the model fits the data well (Hardy & Field 1998 give further explanation and examples). A commonly used alternative test statistic is Pearson's X^2, which has the same asymptotic χ^2-distribution as the deviance.

Several analogues of the r^2 measure commonly used in linear models have been proposed (e.g. Hosmer & Lemeshow 1989), but these do not possess the same statistical meaning and are not as widely used: one could, for example, give the percentage of the deviance explained by each term in the model.

3.4.4 Overdispersion and underdispersion

For a well-fitting model, the residual deviance should be approximately equal to the residual degrees of freedom (i.e. the *residual mean deviance* (residual deviance/residual *df*) should be approximately equal to one and certainly less than about 1.5). When this is not the case, either the model does not adequately describe the variation in the data, or the variation in the data is greater than that under binomial sampling. Either way, the most likely result is that the mean deviance will be greater than one. When the model is thought to be correct (i.e. we believe that all important explanatory terms have been included), but the residual mean deviance is greater than one, the data are said to exhibit *extra-binomial variation*, *super-binomial variation* or *overdispersion* (Chapter 5).

3.4.4.1 Causes of overdispersion

There are two main causes of overdispersion in grouped binary data expressed as proportions: either the model is mis-specified in some way, or there is correlation between the responses (i.e. the sexes). Mis-specification could be due to one of the following: (1) a systematic component of the model has been mis-specified (e.g. important variables have not been measured, important interaction terms have been omitted or an explanatory variable needs to be transformed); (2) there are one or more outliers in the observed dataset; (3) an inappropriate link function has been chosen (for proportion data, a complementary log-log link may reduce the degree of overdispersion); or (4) the proportions are based on small numbers of individuals (under these circumstances, the chi-square approximation to the distribution of deviance breaks down and hence a large residual mean deviance may not be problematic).

Once these possible explanations have been eliminated, the most likely explanation for overdispersion in binomial data is *correlation between the binary responses*. In essence, this means that there has been a violation of the assumption that the individual binary observations (i.e. the individual organisms) making up the binomial proportions (i.e. the sex ratios) are *independent* of each other. Since individuals are often grouped together within clutches, broods or families, if sex ratios are biased in any way the individual binary data points (offspring sexes) will be positively correlated, leading to sex ratios that are more variable than they would have been under the assumption that the sexes were distributed binomially. As a consequence, the residual mean deviance will be greater than unity. Overdispersion can be generated not just by inter-clutch variation in sex ratio, but also by any factor that leads to individual binary responses being nonindependent. Overdispersion is also common in mortality data as groups of individuals may tend to survive or die collectively (Chapter 5); Jim Morrison's 'no one here gets out alive'!

3.4.4.2 Correcting overdispersion

Since overdispersion simply means that the variance is greater than that expected under the binomial expectation, the simplest solution is to assume that the variance is not equal to $np(1-p)$, as assumed by the binomial probability distribution, but is *proportional* to it and equal to $np(1-p)s$, where s is an unknown scaling factor variously referred to as the *scale parameter*, *dispersion parameter* or *heterogeneity factor*. We can estimate s by dividing the Pearson's X^2 value (or simply the *residual deviance* for the full model) by the residual degrees of freedom. We can then use this estimate of s (usually termed the *empirical scale parameter*) to compare the *scaled deviances* for terms in the model using F-tests, rather than χ^2 tests (in exactly the same way as we would for a conventional linear model). Applying an empirical scale parameter does not affect parameter estimates, but it does inflate their standard errors (which are multiplied by a factor $\sqrt{s}$); thus, type II errors are more likely (and type I errors less likely). This approximation works well, and is the standard method used by most ecologists (indeed, some ecologists would advocate the use of F-tests rather than χ^2 tests for all GLMs with binomial errors whenever there is any overdispersion, especially when sample sizes are small).

Models using empirical scale parameters are prone to inaccuracies when sample sizes (denominators) vary dramatically between proportions. Williams (1982) developed an alternative method that allows for unequal sample sizes by applying an additional weighting function to the data; this method is now known as *Williams' procedure*. A number of statistics packages have the facility to implement Williams' procedure, including *GLIM* and *Genstat*, but not *S-Plus*. F-tests (or t-tests) should be employed to evaluate the significance of variables after using Williams' procedure.

A further method is *quasi-likelihood estimation*. This allows estimation of regression relationships without fully knowing the error distribution of the response variable. Thus, instead of providing an error distribution and link function, one provides a link function and a variance function. For example, perhaps the logit link transformation linearizes the response correctly, but the variance appears to be a linear function of the mean; under these circumstances, both attributes could be incorporated into a quasi-likelihood model. Quasi-likelihood also allows one to estimate the

scale parameter in under- and overdispersed regression models. For example, in sex ratio analyses, we can estimate the degree of overdispersion in a logistic regression model by supplying the appropriate link and variance functions for the logistic model and determining the significance of parameter estimates using *F*-tests.

Yet another method of dealing with overdispersion is to use *generalized linear mixed models* (GLMMs). One of the problems with GLMs is that they allow only one error term; all effects other than the residual error at the lowest level of the data are assumed fixed (McCullagh & Nelder 1989). Therefore, for parameter estimation purposes, each offspring in the sex ratio is treated as an independent data point. However, if variation between 'clusters' (e.g. broods) does not follow the binomial expectation (e.g. due to sex ratio manipulation), then there will be overdispersion and these estimates will be biased. While we might use an empirical scale parameter, Williams' correction or quasi-likelihood to deal with this (see above), an alternative method would be to introduce a second random effect (e.g. the identity of the brood) to deal with the 'between-cluster' variation in sex ratio, in addition to the 'within-cluster' variation, i.e. to use GLMMs (Krackow & Tkadlec 2001). These models will give the same parameter estimates for the fixed effects as the conventional GLMs, but their standard errors will be inflated if the random effect (e.g. nest identity) is influential. Currently, GLMMs are possible in only a few statistical packages (e.g. *Genstat*, but not *S-Plus* or *GLIM*), but their use and availability are likely to grow. For recent examples of the use of GLMMs in sex ratio analyses, see Kruuk *et al.* (1999) and section 3.5.4. Further details of dealing with overdispersion are given by Collett (1991) and Crawley (1993). It is important to emphasize, however, that if overdispersion is very large, then this indicates a badly fitting model and it might be that a different approach would reflect the biology of the system better (e.g. log-linear modelling of the number of males in the clutch).

3.4.4.3 Overdispersion in binary data

Since the deviance for (ungrouped) binary data does not exhibit a χ^2 distribution, its magnitude depends solely on the value of the fitted probabilities. Therefore, large values of residual mean deviance for binary data cannot be taken to indicate overdispersion. Overdispersion may still occur in binary data, but it will not be possible to detect it from the value of the residual mean deviance and it can be modelled only by including a random effect in the model (Collett 1991).

3.4.4.4 Underdispersion

Underdispersion occurs when the variance of a binomial response variable is less than that for the binomial distribution and may be produced when the individual binary observations are *negatively* correlated. Although underdispersion is rare in sex ratio analyses of vertebrates and most invertebrates, it is common among haplodiploid insects and mites (Chapter 5). Despite this, it has yet to receive much attention from statisticians (but see Podlich *et al.* unpublished manuscript). Another reason is that the costs of ignoring underdispersion appear to be relatively small, as it simply leads to conservative tests, i.e. tests in which the chance of a type I error is not increased. On the other hand, ignoring underdispersion reduces the statistical power of the test and hence increases the chances of making type II errors. In the absence of alternative methods, rescaling the data in the same way as for overdispersed data is recommended (Gordon K. Smyth pers. comm.). For example, Hardy and Mayhew (1998) found a significant negative relationship between mean sex ratio and mean clutch size across 26 species of bethylid wasps using classical regression of arcsine-transformed sex ratio data. When we analysed the same data using logistic regression, the relationship appeared to be nonsignificant ($\chi^2_1 = 0.86$, $P > 0.1$). However, the model exhibited considerable underdispersion (heterogeneity factor $= 0.178$) and when rescaling was applied, using Pearson's X^2, it transpired that the relationship was indeed significant ($F_{1,24} = 27.0$, $P < 0.001$). Analysis of species-mean data is discussed in Chapter 6.

3.4.5 Model simplification

The aim of statistical modelling is to produce a model that fits the data well while also being as simple as possible: this is known as the

Table 3.4 The sequence of steps in model simplification (after Crawley 1993)

Step	Procedure	Explanation
1	Fit the maximal model	Fit all the factors, interactions and covariates of interest Note the residual deviance If you are using Poisson or binomial errors, check for overdispersion and rescale if necessary
2	Begin model simplification	Inspect the parameter estimates Remove the least significant terms first, starting with the highest order interactions, progressing on to lower order interaction terms and then main effects Remember that main effects that figure in significant interactions should not be deleted
3	If the deletion causes an insignificant increase in deviance	Leave that term out of the model Inspect the parameter values again Remove the least significant term remaining
4	If the deletion causes a significant increase in deviance	Put the term back in the model These are the statistically significant terms as assessed by deletion from the maximal model
5	Keep removing terms from the model	Repeat steps 3 or 4 until the model contains nothing but significant terms This is the minimal adequate model If none of the parameters is significant, then the minimal adequate model is the null model

principle of parsimony or *Occam's razor*; in other words, a model that does not contain any redundant parameters or factor levels. Fitting GLMs is a *journey of exploration*! Often, there is no single best model; several models may adequately fit the data and different modelling procedures may yield very different solutions. But remember that at all times *biology should drive your choice of models*. Indeed, Hosmer and Lemeshow (1989) have argued that 'successful modelling of a complex data set is part science, part statistical methods, and part experience and common sense'.

The first step in the model simplification process is to fit a *maximal model* that contains all of the factors, covariates and interaction terms that might be important in the analysis (Table 3.3). Then, via a series of *step-wise deletion tests* (section 3.4.5.1), any nonsignificant explanatory variables, factors and interaction terms are removed, starting with the highest order terms (e.g. three-way interactions). Once the number of terms in the model has been reduced such that no more can be removed without reducing the model's explanatory powers (i.e. causing a statistically significant reduction in the amount of variation explained), and none can be replaced that increase the model's explanatory powers, it may be possible to simplify the model still further by grouping together factor levels that do not differ significantly from one another (aggregation) and amalgamate explanatory variables that have similar parameter values (as long as such simplifications make good biological sense). The resultant model is the *minimal model* (Tables 3.3 and 3.4).

Crawley's (1993, 2002) books contain whole chapters on model simplification and it is well worth reading one of these prior to embarking on any GLM exercise. His views on the sequence of steps in the model simplification process are summarized in Table 3.4 but, as Crawley himself is at pains to point out, there are no hard and fast rules.

3.4.5.1 Determining the significance of individual terms in the model

Step-wise deletion tests are χ^2 tests (or *F*-tests) that assess the significance of the increase in deviance that results when a given term is removed from the current model. For example, imagine we have two hierarchical models (i.e. two models for which one of the models contains all of the terms of the other model, plus one or more additional terms) - model 1: $y = a + bx_1 + cx_2 + dx_3$; and model 2: $y = a + bx_1 + cx_2$, which differ in that model 2 does not contain the dx_3 term. To test the significance of the parameter d, we determine the likelihoods for model 1 (l_1) and for model 2 (l_2), and calculate the (change in) *deviance* for the comparison of the two models $[-2\ \ln(l_2/l_1)]$, which can then be compared to χ^2, with degrees of freedom equal to the difference in the number of terms in the two models. Here, $P < 0.05$ indicates that the variable was making a significant contribution to the fit of the model and hence should generally be retained. However, for very large sample sizes, or where there are many higher-order interaction terms, statistically significant results may be generated even though the effect sizes are small. In these instances, it may be prudent to increase the critical probability level for retention in the model. For example, a good rule of thumb is that the acceptance probability is set at 5% ($P < 0.05$) for main effects, 1% for two-way interactions, 0.5% for three-way interactions, and so on (MJ Crawley pers. comm.).

A less rigorous method of evaluating the significance of a variable in a statistical model is the *Wald-test*, which tests whether the regression coefficient is significantly different from zero by comparing the estimated coefficient to its standard error. In practice, the Wald-test is usually used as a guide to the sequence in which variables are removed from the model, and the amount of deviance the variable explains is used as the final criterion of its significance.

3.4.6 Model checking

Once the minimal model has been obtained it can be checked using a number of regression diagnostics, discussed in detail by, for example, Hosmer and Lemeshow (1989) and Crawley (1993, 2001). These include assessing the overall fit of the model (section 3.4.3) and producing diagnostic plots. For example, we need to assess whether the standardized residuals exhibit any trends with respect to the explanatory variables or fitted values (Figure 3.3a), and whether the standardized residuals are normally distributed (Figure 3.3b). We use *standardized* residuals when the error distribution is binomial (or Poisson or gamma) because the variance changes with the mean (Crawley 1993). Examples of both a 'residuals plot' and a 'normality plot' are shown in Figure 3.3. A lack of pattern in the residuals plot indicates a well-specified model, while the normality plot should generate a reasonably straight line when the model provides a good fit to the data. However, while these plots are good for detecting extreme observations deviating from a general trend, extreme caution should be exercised in over-interpreting them. This is particularly true for binary data, because all of the points on the residuals plot lie on one of two curves depending on whether the response is 0 or 1. Diagnostic plots are produced as standard in *S-Plus* and some other statistical packages, and Crawley (1993, p. 288) provides a macro for generating them in *GLIM*; as well as an example for binomial data.

3.5 Logistic analysis of sex ratio data

Having set the GLM scene, we now examine the GLM modelling process as it applies to proportion data in general, and sex ratios in particular. Logistic regression is the term used to describe GLMs in which the error distribution is assumed to be binomial and a logit link function is applied (section 3.4). Many statistical packages now include logistic regression as a special modelling procedure, even if they also have a generic GLM function (e.g. *S-Plus*) or have no other GLM functions (e.g. *Minitab*) (see Apendix 3.1). Logistic regression can be used to model both binary and binomial (grouped binary) data. The statistical methodologies for analysing these two data types are essentially the same. We begin with the

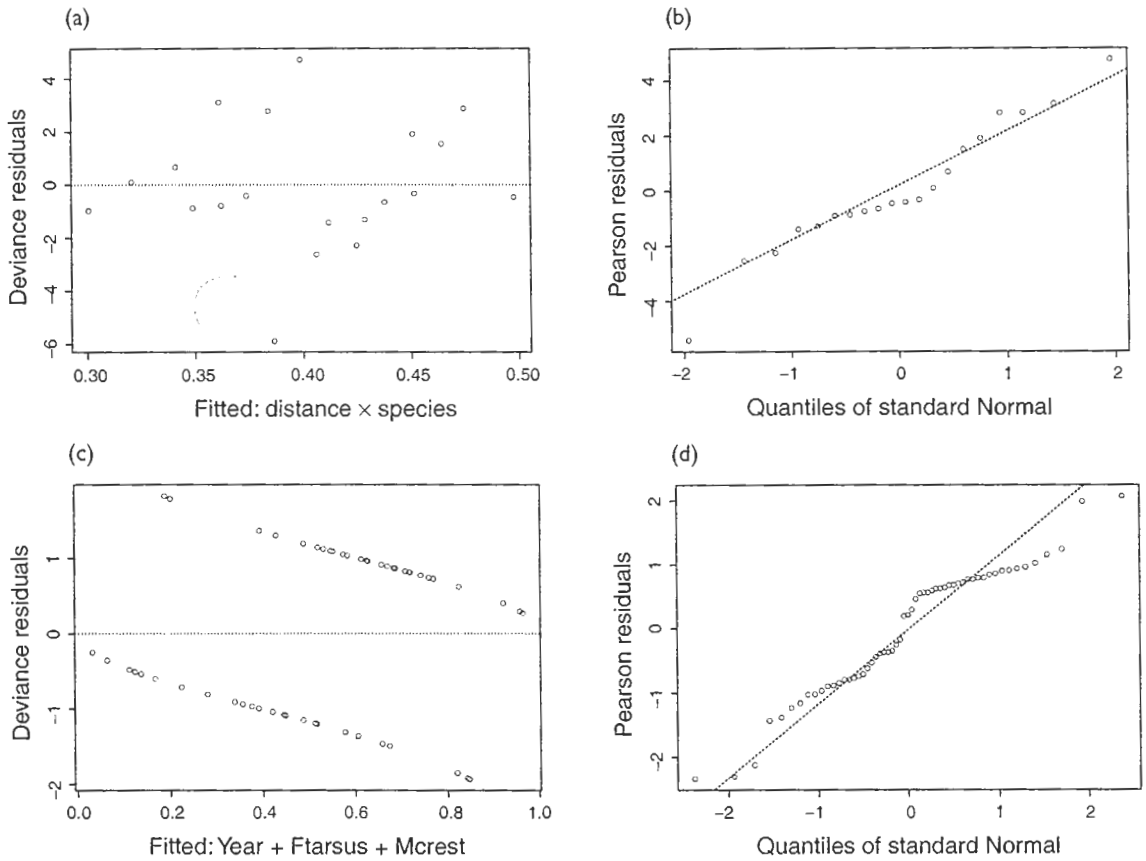

Fig 3.3 Model-checking plots from a logistic regression model. The data are plots derived from a logistic regression model fitted to binomial data from Example 1 (panels a & b) and binary data from Example 2 (panels c & d). In (a) and (c), the deviance residuals are plotted against the fitted values; we refer to these as *residuals plots*. In (b) and (d), the ordered Pearson residuals are plotted against quantiles of the standard Normal distribution; we refer to these as *normality plots*. When the data are binomial (grouped binary), a random scatter of points around zero indicates a well-fitting model, as shown in (a). When the data are (ungrouped) binary, residuals plots are not very useful because all of the points lie on one of two curves depending on whether the response is 0 or 1 (c). For both binomial (b) and binary (d) data, deviation from the line of unity on a normality plot may indicate a poorly fitting model; both of these models appear to fit reasonably well.

analysis of binomial data (section 3.5.1) and follow this up with the analysis of binary data and highlight where the differences lie (section 3.5.2). We give a worked example of analysis of avian sex ratios (section 3.5.3) and discuss a case history of analyses of mammalian sex ratios (section 3.5.4). Logistic analyses of social insect sex ratios are discussed in Chapter 4.

3.5.1 Analysis of proportions

Because sex ratios are proportions, and involve dividing one integer by another, important information about the size of the sample from which they were calculated is lost. This is one of the main problems with traditional (non-weighted) methods that rely on classical regression or nonparametric statistics (section 3.5.3). When proportions are modelled by logistic regression, this information is regained because information about ratios (e.g. number of males versus number of females or number of successes versus number of failures) *and* sample sizes (i.e. the magnitude of the *binomial denominator*) is included.

In most statistical packages, the data (e.g. sex ratio) are included in the model as two vectors: one describing the ratio, and the other the sample size, or (as in GLIM) one describing the

numerator (e.g. males) and the other the sample size (i.e. denominator = males + females). In others (e.g. *S-Plus*), the two vectors may be bound together (using the *cbind* command) and represent the raw data that combine to make the ratio (e.g. number of males and number of females). In all cases the method of analysis involves performing a *weighted* regression using the individual sample sizes as weights, and a *logit link* function to linearize the model (section 3.4.2.3).

3.5.1.1 Example 1: Fish sex ratios revisited

We begin by going back to Example 1. You will remember that we wanted to determine whether pollution from a known source resulted in biased sex ratios in two species of fish living in an Australian creek. Having conducted a visual inspection of the data, we can proceed to fitting our maximal model, which in this case includes just three terms, `Distance`, `Species` and the `Distance:Species` interaction. In *S-Plus*, we can specify this model in one of two ways:

```
(1) model1_glm (SexRatio~Distance*Species, family=binomial,
      weights=SampleSize, data=Example1)
(2) model1_glm (cbind (NumMales,NumFemales) ~Distance*Species,
      family=binomial, data=Example1)
```

In both cases, we tell *S-Plus* that we are creating a generalized linear model (`glm`) with binomial errors and logit link function (`family=binomial`; logit link is the default option), and the data we want to analyse are in the dataset called Example 1 (`data=Example1`). The difference lies in how we tell *S-Plus* that we have proportion data with a known denominator. In (1), we give *S-Plus* a term for the proportion of males (`SexRatio`) and a term for the denominator of the ratio (`weights=SampleSize`), whereas in (2) we give it the two vectors that together indicate the magnitude of the denominator (`cbind (NumMales, NumFemales)`; `cbind` simply 'binds' these two vectors together, such that the number of males in the sample is paired to the number of females from the same sample; the denominator = `NumMales+NumFemales=SampleSize`).

In both cases, we get the following output (the diagnostic plots for this model are shown in Figure 3.3a,b):

```
Coefficients          Value     Std. Error   t value
(Intercept)         -0.8069     0.1779      -4.5343
Distance             0.0007     0.0002       2.8799
Species             -0.1304     0.1779      -0.7328
Distance:Species     0.0001     0.0002       0.7791

(Dispersion Parameter for Binomial family taken to be 1)
  Null Deviance: 118.0099 on 19 degrees of freedom
Residual Deviance: 107.7797 on 16 degrees of freedom

Terms added sequentially (first to last)
Term              Df   Deviance   Resid.Df   Resid.Dev   Pr(Chi)
NULL                                19        118.0099
Distance          1    9.6185       18        108.3914    0.0019
Species           1    0.0023       17        108.3890    0.9612
Distance:Species  1    0.6092       16        107.7797    0.4350
```

The first table in this output tells us, for each of the coefficients, the value of the parameter estimate (`Value`), its standard error (`Std. Error`) and a *t* value comparing the estimate against zero (`t value`). The second table tells us the change in the number of degrees of freedom (`Df`) and change in deviance (`Deviance`) associated with

the sequential inclusion of each of the terms (Term) in the model and the statistical significance of the change in deviance, as determined by chi-square tests (Pr(Chi)). The other values in this table indicate the sequential reduction in the residual degrees of freedom (Resid.Df), and residual deviance (Resid.Dev). Sandwiched between these two tables are three important lines. These tell us that the statistical analysis of this model assumes that the dispersion parameter is taken to be 1; in other words that there is no overdispersion (section 3.4.4). Is this true? We can get a rough idea of this by dividing the residual deviance by the residual degrees of freedom (Resid.Dev/Resid.Df = 107.8/16 = 6.73). Clearly, there is massive overdispersion, whereas our model currently assumes that there is none (i.e. that Resid.Dev/Resid.Df ~ 1). Having checked that we have included all possible terms in our model, and that it has not been mis-specified in any way (e.g. by omitting an important interaction term), that we have not ignored any outliers, and that we have the correct link function (section 3.4.2.3), it seems likely that we have genuine overdispersion. This is perhaps not too much of a surprise given that these data are not real, but we shall not let that worry us at this stage. To proceed as we would do with real data, we need to employ an empirical scale parameter ($s = 6.73$). In *S-Plus*, we do this simply by testing the significance of terms in the model using *F*-tests, rather than *chi-square* tests. We begin the process of step-wise deletion by testing the significance of the interaction term. This produces the following output (which has been adapted from *S-Plus* to make it clearer):

```
model2_update(model1,~.-Distance:Species)
anova(model1,model2,test="F")
```

Resid.Df	Resid.Dev	ΔTerms	ΔDf	ΔDeviance	F-value	P(F)
17	108.3672	-Distance:Species	−1	−0.6132	0.0955	0.7612

The first command simply removes the interaction term from our maximal model (model 1). The second command asks *S-Plus* to examine the difference between the amount of variation explained by models 1 and 2 (with and without the interaction term) using *F*-tests. In the table, Resid.Df and Resid.Dev are the residual deviance and residual degrees of freedom, respectively, for the model that excludes the terms indicated by ΔTerms (Δ, 'delta', simply means 'change in'). The table tells us that the process of removing the Distance:Species interaction generates a final model that has a deviance of 108.36 and 17 degrees of freedom, and results in a change in deviance of 0.6132 and 1 degree of freedom. But, remember that we are no longer interested in deviances, because our data are overdispersed. We therefore need to concentrate on the *F*-test. This indicates that removing the interaction term from the model does not reduce the amount of variation explained by our model ($F_{1,17} = 0.0955$, $P = 0.7612$). If it did significantly reduce it, then our current model would also be the minimal model (Table 3.3) and the modelling process would be complete for this particular example. However, as it doesn't, we need to go on to test each of the main effects in turn, starting with the term with the lowest *t* value. In *S-Plus*, this is how we would do it:

```
Model3_update (model2,~.-Species)
Model4_update (model2,~.-Distance)
anova (model2,model3,test="F")
anova(model2,model4,test="F")
```

Resid.Df	Resid.Dev	ΔTerms	ΔDf	ΔDeviance	*F*-value	*P*(*F*)
18	108.3697	-Species	−1	−0.0025	0.0004	0.9839 ns
18	116.1776	-Distance	−1	−7.8104	1.2856	0.2726 ns

Thus, although Distance appears to have a bigger effect on the fit of the model than

Table 3.5 | Summary of analyses of fish sex ratios (Example 1)

Model	Test	Type of model	Species	*P*-value (`Distance`)
1	Spearman's rank order correlation	Nonparametric	Shirazfish	$P > 0.16$ ns
			Merlotfish	$P = 0.063$+
			Combined *P*	$P = 0.036$*
2	Pearson's product-moment correlation (arcsine-transformed data)	Parametric – normal errors	Shirazfish	$P > 0.14$ ns
			Merlotfish	$P = 0.057$+
			Combined *P*	$P = 0.029$*
3	General linear model (unweighted, untransformed data)	Parametric – normal errors	Both	$P = 0.013$*
4	General linear model (unweighted, arcsine-transformed data)	Parametric – normal errors	Both	$P = 0.012$*
5	General linear model (unweighted, logit-transformed data)	Parametric – normal errors	Both	$P = 0.012$*
6	General linear model (weighted, arcsine-transformed data)	Parametric – normal errors	Both	$P > 0.16$ ns
7	Generalized linear model (weighted, untransformed data)	Parametric – binomial errors	Both	$P > 0.21$ ns

ns = $P > 0.1$, + = $0.05 < P < 0.1$, * = $P < 0.05$.

`Species`, neither term is statistically significant (`Species`: $F_{1,18} = 0.0004$, $P = 0.9839$; `Species`: $F_{1,18} = 1.2856$, $P = 0.2726$). This suggests that there is no consistent effect of pollution on sex ratio in this population. Just to be sure, we should try adding terms back into the model, starting with `Distance`. When `Distance` alone is added to the model, no significant variation in sex ratio is explained ($F_{1,18} = 1.675$, $P = 0.21$), even if we employ quasi-likelihood estimation to obtain a better level of compensation for overdispersion ($F_{1,18} = 1.726$, $P = 0.21$).

3.5.1.1.1 SUMMARY OF EXAMPLE 1

In summary, if we compare the performance of the different tests (Table 3.5), we see that the non-parametric tests (Spearman's correlation) gave lower significance values for `Distance` than the equivalent parametric tests (Pearson's correlation) (cf. models 1 and 2) (section 3.3.1.2). This is almost certainly due to this test's lack of power. Using a general linear model (in effect, an ANCOVA), we were able to combine a factor and a covariate within a single model, and this improved the significance level associated with `Distance`, regardless of whether we transformed our sex ratio data or not (cf. model 2 and models 3, 4 and 5) (section 3.3.2). However, when we added a weighting factor to our model, to control for differences in sample size within our dataset, `Distance` disappeared as a significant term in the model (cf. models 5 and 6) (section 3.3.3). Applying a GLM with binomial errors and logit link function yielded similar results, and confirmed that there was no significant change in sex ratio with distance from the pollution source (section 3.5.1.1). The similarity between the results of models 6 and 7 is probably due to the overriding importance of sample size effects in this analysis (i.e. power, rather than the lack of fit of the data to the normal distribution). Careful examination of Table 3.5 indicates that

although the single-species nonparametric tests gave the correct result (i.e. no effect of `Distance` on sex ratio) it appears to have given it for the wrong reason (i.e. due to lack of power)!

At this point, it is worth emphasizing that although our comparison of the different methods has focused on the statistical significance of the result (i.e. the P value), as biologists we are usually more interested in the biological significance of the result rather than its statistical significance (though journal editors may sometimes disagree!). If sample sizes are large enough, then even a 1% difference between treatment groups will be statistically significant (this is why pollsters question such large numbers of people in the run-up to elections). Thus, it is not sufficient to consider just the statistical significance of any trends in our data, but also their magnitude. Thus, in Example 1, the equation for the (non-significant) logistic regression was

$$\text{Logit}(\texttt{Sex ratio}) = -0.7504 + 0.0006808 \times \texttt{Distance}.$$

Thus, back on the *original scale*, the sigmoidal relationship between sex ratio and distance from the pollution source is described as follows

$$\texttt{Sex ratio} = \frac{e^{(-0.7504 + 0.0006808 \times \texttt{Distance})}}{1 + e^{(-0.7504 + 0.0006808 \times \texttt{Distance})}}.$$

Thus, the sex ratio was predicted to vary from 0.33 at the source (100 m) to 0.48 at the furthest distance from the source (1000 m). Since this a fairly large increase in the proportion of males (45%) over a relatively short distance, it would be premature to dismiss pollution as a correlate of sex ratio variation at this stage and we might want to gather a new, larger dataset that will increase the power of our analysis.

3.5.2 Analysis of binary data

Often, sex ratio data are best analysed in the form of binary responses. The analysis of binary data using GLMs is exactly the same as for binomial (grouped binary) proportions, except that we do not include any weighting factor because each 'sex ratio' (0 or 1) represents a single individual and we cannot detect or correct for overdispersion (section 3.4.4.3). Effectively, we assume that each data point comes from a binomial trial in which the sample size (n) is equal to 1. In other words, the data are assumed to come from a special, abbreviated form of the binomial distribution, known as the *Bernoulli distribution* (Collett 1991). Whether it is worth analysing data in this format (rather than as sex ratios based on lumping together individuals from similar groupings, e.g. nests or sampling points) is largely dependent on whether each individual in the analysis has unique explanatory variables associated with it (e.g. an individual weight or colour, or individuals are produced one at a time by parents, i.e. brood size = 1, etc.). If it does, then the data are best analysed in binary form; if not then there is little to be gained and the data can be lumped without loss of information.

3.5.2.1 Example 2: Crest size in Crested Auklets

To address this issue, we examine the relationship between chick sex and paternal crest size in the Crested Auklet (Box 3.7 gives background information). Hunter *et al.* (in prep.) collected

Box 3.7 Example 2: Crest size in Crested Auklets

Data on the relationship between crest size and chick sex in the Crested Auklet (*Aethia cristatella*) was compiled by Fiona Hunter and colleagues (Hunter *et al.* in prep.). These small seabirds breed in coastal colonies around the Bering Sea, nest in crevices and produce just one chick each year. The adults are socially monogamous and both sexes prefer mates with a large crest (a sexual ornament sprouting just above the beak). Hunter *et al.* wanted to know whether females were more likely to produce male chicks when they were paired to males with longer crests. Since there is mutual sexual selection in this species, Hunter *et al.* predicted that, provided crest length was heritable, females would produce sons if they were paired to long-crested males, and daughters if they were paired to short-crested males.

Table B3.7a Sex ratios in Crested Auklets across three years

Year	Number of female chicks	Number of male chicks	Total number of chicks
1993	11	15	26
1994	3	3	6
1995	13	12	25
Total	27	30	57

Data were collected from 57 breeding pairs over three years: 1993, 1994 and 1995 (Table B3.7a). In each year, Hunter *et al.* recorded the sex and mass of the chick each pair produced, plus the body mass, tarsus (leg) length and crest length for the male and female parents (referred to as the sire and dam, respectively). Table B3.7b shows data for 1993 only.

Table B3.7b Chick sex and morphometric data in Crested Auklets in 1993

		Sire			Dam		
Pair number	Chick sex	Mass (g)	Tarsus (mm)	Crest (mm)	Mass (g)	Tarsus (mm)	Crest (mm)
1	M	293	29.7	43.1	255	26.2	37.7
2	F	278	28.2	40.6	271	26.6	36.2
3	M	258	27.8	42.6	234	27.4	52.0
4	M	289	28.2	41.2	256	27.2	45.5
5	F	276	28.0	38.8	235	27.4	44.4
6	F	256	28.0	36.2	270	28.8	39.2
7	M	306	29.4	39.4	254	28.3	35.0
8	F	248	25.8	31.3	244	27.3	39.5
9	M	254	28.8	38.9	269	28.8	36.2
10	M	264	30.3	49.7	272	29.4	42.0
11	M	267	28.6	36.2	271	29.7	35.2
12	M	309	28.4	40.4	265	28.4	36.4
13	M	308	28.9	44.8	256	27.9	35.4
14	F	271	29.6	34.6	264	28.3	42.6
15	M	248	28.4	38.3	261	29.1	40.2
16	F	282	26.1	36.6	239	27.3	39.2
17	M	271	28.3	36.3	255	29.0	36.5
18	F	241	29.3	35.9	249	28.5	38.0
19	M	262	27.5	33.6	257	29.5	34.1
20	M	244	28.9	40.4	272	28.8	38.3
21	F	274	27.7	39.5	236	27.2	37.0
22	F	258	28.4	41.5	283	29.4	37.4
23	F	275	26.6	41.0	242	26.3	42.4
24	F	277	27.3	30.6	252	28.0	36.0
25	M	281	28.8	38.4	270	28.1	38.0
26	M	271	28.1	43.3	–	28.2	38.8

these data to find out whether females were more likely to produce male chicks when they were paired to males with longer crests. Before addressing this issue (section 3.5.2.1.4) we ask a simpler question, namely: does sex ratio vary between years? There are several ways that we can address this question.

To analyse the Example 2 data, we need to organize them so that there is a single dependent variable (`Count` = number of chicks in a given category), and two factors (each with two levels) corresponding to `Year` and `Sex`, and then implement a GLM with Poisson errors and log-link function. In *S-Plus*, the resulting model is:

```
model4_glm(Count~Year+Sex, family=poisson, data=Example2a)

Terms added sequentially (first to last)
      Df  Deviance  Resid.Df  Resid.Dev  Pr(Chi)
NULL                     3     1.4031
Year   1   0.8618        2     0.5413     0.3532
Sex    1   0.1579        1     0.3833     0.6910
```

The significance of the model is tested by comparing its residual deviance (0.3833) with the tabulated χ^2 statistic with 1 degree of freedom (3.841). Since the calculated χ^2 statistic is lower than the critical value in the tables ($\chi^2_1 = 0.3833$, $P > 0.53$), we cannot reject our null hypothesis that the two sexes are distributed randomly across years (i.e. that sex ratio varies between years). Note that if we had a more complicated model, with more factors, we would be better off starting the analysis by constructing a saturated model (Table 3.3), so that we end up with zero deviance and zero degrees of freedom. This would allow us to determine that we had all possible factors in the model before we began the stepwise deletion process (Table 3.4).

3.5.2.1.1 CONTINGENCY TABLES

Often, the simplest way is to construct a $2 \times n$ *contingency table* and calculate *Pearson's chi-square* to test the null hypothesis that individuals are distributed independently with respect to year and sex. A problem with the data in Table B3.7a is that the sample sizes are rather small for 1994. Therefore, in our analyses we shall combine the 1993 and 1994 data (combining 1994 with 1995 gives similar results). This generates a 2×2 contingency table and a chi-square test gives $\chi^2_1 = 0.38$, $P = 0.54$, suggesting that the sex ratio is similar in all years.

3.5.2.1.2 LOG-LINEAR MODELS

A better method for analysing these types of data (often called the *G-test*) extends the contingency table approach and uses *log-linear models*. These are generalized linear models for modelling Poisson-distributed data (as opposed to binomial data). Like the chi-square test, log-linear models yield a χ^2 statistic. Their great advantage is that they can be readily generalized to analyse datasets that are much more complicated than simple 2×2 contingency tables (e.g. Crawley 1993). Moreover, since log-linear models are GLMs, their relation to the other models we have discussed is more easily appreciated.

3.5.2.1.3 LOGISTIC REGRESSION

A third way of looking at these data is to convert them to *proportions* and analyse them using *logistic regression* (rather than analysing them as counts using log-linear regression). A model in which just the intercept term is fitted yields a residual deviance (0.3833 with 1 *df*) that is equal to that determined by the log-linear model, and again indicates that there is no significant variation in sex ratio between years.

We can perform exactly the same analysis by constructing an unweighted logistic regression model using the *raw* (binary) data (Table B3.7a):

```
Model5_glm(Sex~Year, family=binomial, data=Example2b)
```

In this case, we obtain the following output:

Terms	Df	Deviance	Resid.Df	Resid.Dev	Pr(Chi)
NULL			56	78.8608	
Year	1	0.3833	55	78.4774	0.5358

Again, we find that there is no difference in sex ratio between 1993/94 and 1995 ($\chi^2_1 = 0.3833$, $P = 0.5358$). In fact, log-linear and logistic regression models are exactly equivalent when the response is two level (Aitkin *et al.* 1989, pp 225–255).

3.5.2.1.4 PATERNAL CREST LENGTH AND CHICK SEX

Here we address the question of whether there is an association between the sex of chick produced by female Crested Auklets and paternal crest length. The usual first step of plotting the data for visual inspection is difficult when the data are binary because the dependent variable has just two states: male and female. In some instances, summarizing the data with respect to sex can be helpful, but in others (particularly more complex models) it is not. In Example 2, the average crest length of males that sired daughters was 38.77 ± 1.18 mm (s.e.), whereas the average for males siring sons was 42.51 ± 0.99 mm ($t = -2.4543$, $df = 55$, $P = 0.0173$). Thus, males siring sons have longer crests than those siring daughters. The focus of the analysis is, however, on the factors that determine offspring sex, and so the dependent variable is chick sex, rather than male crest length.

One way we could address this question would be to perform a simple logistic regression, with chick sex (Sex) as the dependent variable and male crest length (Mcrest) as the only explanatory variable

```
Model6_glm(Sex~Mcrest, family=binomial, data=Example2b)
```

Terms	Df	Deviance	Resid.Df	Resid.Dev	Pr(Chi)
NULL			56	78.8608	
Mcrest	1	6.0653	55	72.7954	0.0137 *

This appears to confirm that the proportion of sons produced increases with increasing paternal crest length; the relationship is described by the following linear equation (on the *logit* scale):

$$\text{Logit}(\text{Sex}) = -4.8992 + 0.1231 \times \text{Mcrest}.$$

Thus, back on the *original scale*, the sigmoidal relationship between offspring sex ratio and paternal crest length is described as follows

$$\text{Sex} = p = \frac{e^{(-4.899 + 0.123 \times Mcrest)}}{1 + e^{(-4.899 + 0.123 \times Mcrest)}}.$$

Of course, it is possible that this relationship is spurious, generated by some third factor. For example, perhaps good-quality females produce sons and also find good-quality mates with long crests. Alternatively, perhaps, females produce more sons in 'good years' and males produce longer crests in 'good years', leading to a positive correlation between sex ratio and male crest length across years, which is not present within years. Although we cannot examine all possible confounding variables, we can determine whether any of the other variables we measured are important. Hunter *et al.* (in prep) measured a variety of morphometric characteristics in addition to male crest size, including tarsus length and body mass, and they did this for both sexes (Table B3.7b). We also know in which year the measurements were made. Therefore, we are in a position to answer our main question while testing for additional factors that might be either accentuating or masking the relationship between paternal crest length and chick sex.

The first problem is which model to begin with. There are seven possible explanatory variables: one factor and six covariates (for a

reminder about the difference between a factor and a covariate, refer back to Box 3.1). This means that we have 13 main effects, if we include the six possible quadratic terms. If we also fit all 78 pairwise interactions ($12 + 11 + 10 \ldots + 2 + 1$), this means that we have 91 parameters to estimate, yet only 56 data points! In these circumstances, it is less obvious what the correct procedure for model simplification is (section 3.4.5). Clearly, a compromise is needed, and this is where the art of statistical modelling comes into its own (and where individual modellers' opinions may differ). The trick is to start at a sensible point and then go back to test those terms that were initially ignored. Crawley (1993) recommends including not more than $n/3$ parameters in an initial model. Thus, in this example, no more than $56/3 = 19$ terms. Starting with the seven main effects plus six quadratic terms leaves room for just six interaction terms. An alternative starting point might be seven main effects plus 12 interaction terms. There are no hard and fast rules, but we started with the following model

```
model7_glm(Sex~Year* (Mcrest+Mmass+Mtarsus+Fmass+Ftarsus+Fcrest),
      family=binomial, data=Example2)
```

This was based on the idea that we could only reasonably allow one interaction term in the initial model and we felt that the most important interaction terms were likely to involve 'year', but we could easily have chosen to start with interactions involving male crest length or the quadratic terms. Even with just a single interaction term, our model includes 14 terms! Once we had removed all terms that did not contribute significantly to model fit (stepwise deletion tests), we tried to add more terms, including terms that had previously been rejected and high-order interaction terms. In practice, this involved going through each of the seven main effects, one by one, and testing for inclusion all interaction terms involving that effect. It is important to act systematically. We obtained the following minimal model, based on a series of stepwise-deletion tests

```
Model8_glm(Sex~Year+Ftarsus+Mcrest,family=binomial,data=Example2)
```

Resid.Df	Resid.Dev	ΔTerms	ΔDf	ΔDeviance	Pr(Chi)
54	69.3214	−Year	−1	−4.8827	0.0271 *
54	69.4760	−Ftarsus	−1	−5.0373	0.0248 *
54	72.1635	−Mcrest	−1	−7.7247	0.0054 **

Coefficients:

Terms	Value	Std.Error	t-value
(Intercept)	−27.0913	10.1588	−2.6667
Year	−0.7745	0.3773	−2.0527
Mcrest	0.1775	0.0744	2.3831
Ftarsus	0.7014	0.3289	2.1325

Thus, a higher proportion of male offspring were produced in 1995 than in 1993/94, and the proportion of male offspring increased with maternal tarsus length (body size) and paternal crest length. The fitted logistic regression lines are shown in Figure 3.4a,b (diagnostic plots are shown in Figure 3.3c,d).

3.5.3 A worked example: Sex ratio manipulation in zebra finches

In this section, we take an example from the literature to illustrate the advantages of logistic regression over traditional methods. Our aim is not to highlight the weaknesses of the published study (which are not atypical, Box 3.4), but rather to highlight the advantages of the GLM approach.

Our example comes from a study by Becky Kilner (1998) on sex ratio manipulation in zebra

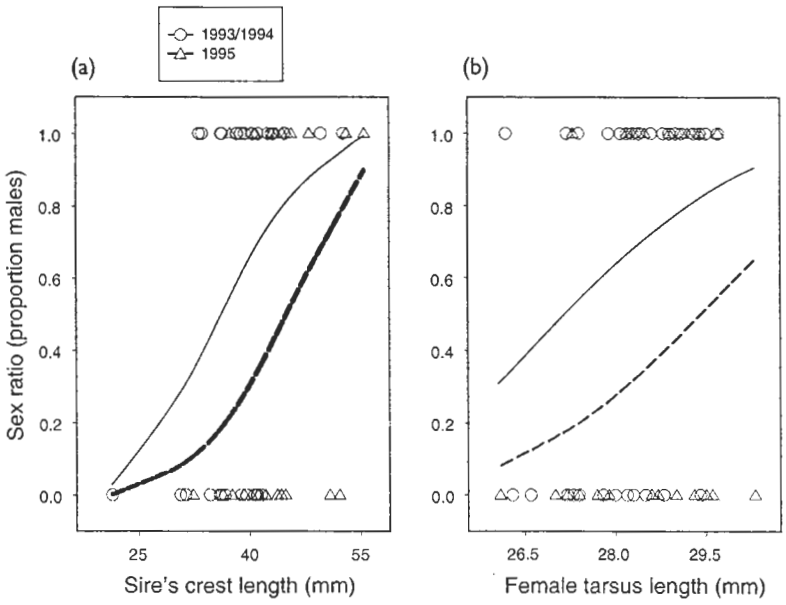

Fig 3.4 Relationships between sex ratio and (a) male crest length and (b) female tarsus length in Crested Auklets. Some of the data used to create these plots are given in Table B3.7b. The curves are the fitted partial logistic regression lines.

finches. The background to this study and the original analysis are given in Box 3.8. Kilner predicted that sex ratios would be more female-biased when food was abundant and more male-biased when food was restricted. Further, since first-hatched chicks tend to attain heavier fledging weights, she reasoned that within broods females would tend to hatch earlier than males. Here, we address the question: how does diet affect the relationship between hatch order and sex ratio?

Our analysis revolves around the data presented in Figure 3.5a. This shows the relationship between the order in which chicks hatch and the proportion of males hatching, for offspring of adult birds fed on either restricted or abundant food. Kilner's analyses (Box 3.8) suggested that sex ratio increased with hatching rank for chicks derived from well-fed pairs, but not for those from pairs given a restricted diet. This conclusion was largely based on a series of nonparametric analyses (outlined in Box 3.8) in which sex ratio was calculated for each of the six hatch ranks separately, resulting in $n = 6$ for each diet. This is despite the fact that the analysis is based on 23 pairs of birds, 42 broods and 162 eggs. This loss of information is particularly important because sample sizes vary considerably across hatch ranks. This point is illustrated in Figure 3.5b, where the size of each

Box 3.8 Example 3: Sex ratio manipulation in zebra finches

This example comes from a study by Kilner (1998) on sex ratio manipulation in the zebra finch (*Taeniopygia guttata*), a small, seed-eating passerine bird that lives throughout the arid and semi-arid zones of Australia and Indonesia. Zebra finches are nomadic and breed opportunistically when there is sufficient food available. There is evidence from other studies that females may manipulate clutch sex ratios in relation to mate quality and food abundance. In wild populations, secondary sex ratios tend to be female-biased when food is abundant, though trends are not consistent between years. In order to test this experimentally, Kilner manipulated the quantity of food available to captive breeding birds and monitored their subsequent primary and secondary sex ratios (here, we restrict our discussion to her analysis of primary sex ratios, i.e. the proportion of males in the brood at hatching). Kilner predicted that sex ratios would be more female-biased when food was abundant and more male-biased when food was restricted. Further, since

first-hatched individuals tend to attain heavier fledging weights, she reasoned that within broods females would tend to hatch earlier than males.

Kilner reared 12 pairs of birds under two regimes of food availability. For their first clutch, all birds were reared on a food-restricted regime in which food was rationed via an electronic hopper. Ten of these pairs then produced a second brood of eggs, again on a restricted food regime. Then, for the third brood ($n = 9$ pairs), the hoppers were removed and food was supplied *ad libitum*. A second group of 11 birds was also established at this point, which laid their first batch of eggs under conditions of abundant food. This was to control for any variation in sex ratio due to the number of broods that a pair had previously reared.

Kilner conducted a number of analyses on these data, but here we concentrate on trying to address one question: how does diet affect the relationship between hatch order and sex ratio?

Original analysis

Kilner performed a series of separate tests designed to answer specific aspects of the main question. For example, using Mann–Whitney *U*-tests she showed that, across all 42 broods, sex ratio was significantly more male-biased when food was restricted than when it was abundant ($P < 0.01$). Using Friedman two-way ANOVAs, she showed that, within the nine pairs of birds for which she had three broods, sex ratios were significantly more male-biased when food was restricted than when it was abundant ($P < 0.05$). Using Wilcoxon signed-ranks tests she showed that, across all 22 'food-restricted' broods, female offspring hatched significantly earlier than male offspring ($P < 0.01$), and a similar relationship was also apparent across the 20 'food-abundant' broods ($P < 0.05$). Using Spearman rank correlations, she showed that within the food-restricted group, the proportion of males hatching increased significantly with increasing hatch order ($P < 0.05$) and a similar, but nonsignificant ($P < 0.1$), trend was apparent in the food-abundant group. Using Wilcoxon signed-ranks tests, she showed that across the six hatch order positions (1–6), the proportion of males hatching was significantly lower when food was abundant ($P < 0.05$). Finally, using the nine pairs for which she had data for three broods, she used Friedman's two-way ANOVA to show that the proportion of males hatching at each rank was significantly lower when food was restricted than when it was abundant ($P < 0.05$). The relationship between diet, hatch order and sex ratio is shown in Figure 3.6a.

While all these tests point to there being a genuine effect of diet and hatch rank on brood sex ratio, this analysis has a number of problems. First, it uses nonparametric tests, which tend to lack power and are susceptible to type II errors (i.e. incorrectly accepting the null hypothesis). Second, while some of the tests underutilize the data, by not weighting sex ratios by clutch or brood size (e.g. the Spearman rank correlations), others appear to be pseudo-replicated (e.g. the Mann–Whitney *U*-tests, where all of the data are lumped together with respect to diet regime, without taking account the identity of the pair). Third, at least six different tests are used, when one test could do the job, as illustrated in section 3.5.3. A problem associated with this approach is that the probability of generating type I errors increases and so Bonferonni corrections generally need to be applied (Rice 1990).

symbol is proportional to the size of the denominator. By plotting the data in this way, it becomes abundantly clear that the original analysis is likely to be unduly influenced by those data points that are based on small sample sizes (e.g. hatch ranks 5 and 6).

We can re-analyse these data using simple logistic regression in which the dependent variable is the proportion of hatchlings that are male and the explanatory variables are `HatchOrder`, `Diet` and their interaction (`HatchOrder:Diet`). The regression is weighted by clutch size (the denominator of the proportion), so making full use of the available data. Of course, as this is logistic regression, we initially assume a binomial error distribution and a logit-link function.

The results of this first step in the analysis is shown below:

Coefficients	Value	Std. Error	t value
(Intercept)	−0.7583	0.4835	−1.5682
HatchOrder	0.4357	0.1672	2.6055
Diet	−0.6223	0.7398	−0.8412
HatchOrder:Diet	0.0508	0.2591	0.1961

(Dispersion Parameter for Binomial family taken to be 1)

Null Deviance: 23.61876 on 11 degrees of freedom
Residual Deviance: 6.32614 on 8 degrees of freedom

Terms	Df	Deviance	Resid. Df	Resid. Dev	$P(\chi^2)$
Null model	11	23.6182			
HatchOrder	1	15.0691	10	8.5487	0.0001
Diet	1	2.1841	9	6.3646	0.1394
HatchOrder:Diet	1	0.0385	8	6.3261	0.8444

Checking for overdispersion by calculating the heterogeneity factor (i.e. `Resid.Dev/Resid.Df` = 6.3261/8 = 0.7907) suggests that there is, in fact, underdispersion, and since it is only slight it is safe to proceed. The output appears to show that there is a significant effect of hatch order on sex ratio, but that diet and the interaction between hatch order and diet is nonsignificant. However, to test this properly, we need to delete each term from the model in turn and determine whether it results in a significant decrease in the proportion of deviance explained (stepwise deletion tests). Since `HatchOrder:Diet` is the only interaction term in the model, and is nonsignificant, we can delete it and test the significance of the two main effects (but, clearly, if there was more than a single interaction term, we would perform stepwise deletion tests for each of the interaction terms as well). This produces the following results:

Resid. Df	Resid. Dev	ΔTerms	ΔDf	ΔDeviance	P (χ^2)
10	8.5487	-Diet	−1	−2.1841	0.1394 ns
10	20.6345	-HatchOrder	−1	−14.2699	0.0002 ***

The loss of `Diet` from the model results in the deviance explained decreasing by a small and nonsignificant amount ($\chi^2_1 = 2.18$, $P > 0.13$). However, the loss of `HatchOrder` from the model results in a highly significant decline in the amount of deviance explained ($\chi^2_1 = 14.27$, $P < 0.001$). So, the only significant explanatory variable is `HatchOrder`, and the analysis of deviance table for this model is shown below (note that when `Diet` is the only term in the model,

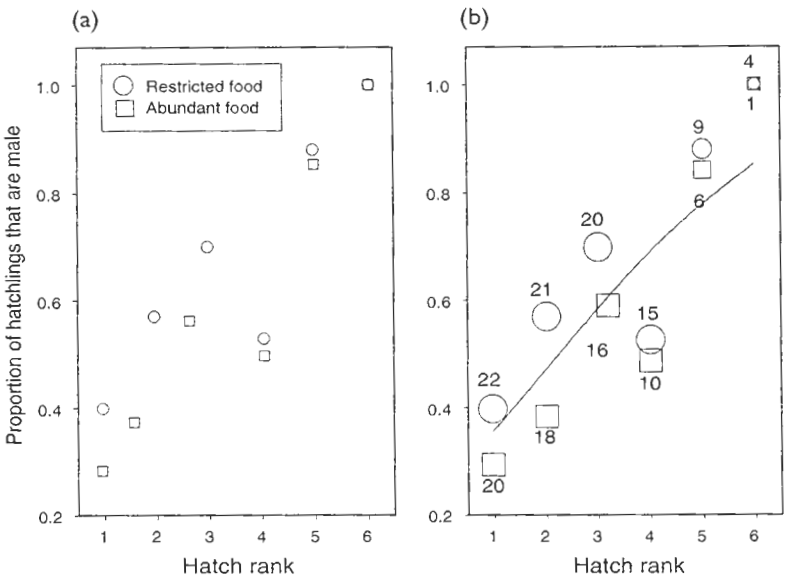

Fig 3.5 Sex ratio at hatching with respect to hatch order, for zebra finch broods reared with abundant (□) and restricted food (○). Data taken from 42 broods, after Kilner (1998, Box 3.8). In (a), the data are shown as they appeared in Kilner (1998); in (b) symbol size is proportional to sample size (which are indicated above or below each symbol) and the line is the fitted logistic regression line to all of the data.

it is marginally nonsignificant: $\chi^2_1 = 2.98$, $P = 0.084$):

Terms	Df	Deviance	Resid. Df	Resid. Dev	P (χ^2)
Null model	11	23.6187			
HatchOrder	1	15.0699	10	8.54877	0.0001 ***

And the summary output for the model (from *S-Plus*) is as follows:

Coefficients	Value	Std. Error	t value
(Intercept)	−1.0523	0.3628	−2.9001
HatchOrder	0.4647	0.1268	3.6632

(Dispersion Parameter for Binomial family taken to be 1)

Null Deviance: 23.618 on 11 degrees of freedom
Residual Deviance: 8.548 on 10 degrees of freedom

This indicates that the intercept for the logistic regression line is significantly different from zero and that there is a significant positive relationship between hatching order and proportion of males in the brood. In other words, by using a weighted analysis of covariance in which we model sex ratio with binomial errors and a logit-link function, it appears that although females tend to hatch before males, this effect is independent of diet, which is nonsignificant.

Remember that the parameter estimates shown here are from the linear predictor (section 3.4.2.2), and so are on a logit scale (logit = $\ln(p/(1-p))$. To back-transform from logits (z) to proportions (p), we apply eq. 3.13 (i.e. $p = e^z/[1+e^z]$). Thus, the predicted sex ratio for first-hatched eggs is $e^{(-1.052+1\times0.465)}/(1+e^{(-1.052+1\times0.465)}) = 0.357$, for second-hatched eggs is $e^{(-1.052+2\times0.465)}/(1+e^{(-1.052+2\times0.465)}) = 0.469$ and for sixth-hatched eggs is $e^{(-1.052+6\times0.465)}/(1+e^{(-1.052+6\times0.465)}) = 0.850$. The logistic regression line derived from this analysis is shown in Figure 3.5b.

The model output also reminds us that this model assumes that the dispersion (scale) parameter is equal to 1 (i.e. on the logit scale, the

variance is independent of the mean). We can re-estimate the scale parameter by dividing the residual deviance by the residual degrees of freedom `8.548/10 = 0.8548`. As this is tolerably close to unity, we need not worry greatly about underdispersion (section 3.4.4.4). One way that we could check the robustness of this result would be by constructing a quasi-likelihood model in which we assumed a logit link, and tested for significance using *F*-tests instead of chi-square tests (section 3.4.4.2). This produces similar results, with hatching order being the only significant term in the model ($F_{1,10} = 20.791$, $P = 0.0010$). However, diet was close to significance in this model ($F_{1,9} = 4.108$, $P = 0.0733$), suggesting that further experiments or analyses may be justified.

3.5.3.1 Further analyses

While our re-analysis makes better use of the available data and has greater power than those conducted by Kilner, it is not the best analysis possible. This is because our analysis weights all 162 offspring equally, regardless of their parents' identity or which brood they came from. Often, these two factors will lead to overdispersion, but the fact that our model is underrather than overdispersed suggests that these effects are not biasing our results systematically. However, Kilner's experimental design allows us to test simultaneously for the independent effects of parentage or brood number on sex ratio. For this analysis, we need to employ generalized linear mixed models (GLMMs, section 3.4.4.2) in which these terms are included as random effects (see Krackow & Tkadlec 2001). When we conduct such an analysis (in *Genstat*, using the *irreml* procedure), and determine the significance of terms in the model using *F*-tests (see Elston 1998), we get results similar to those gained with the GLM: although it is very clear that the `Diet:HatchOrder` interaction is nonsignificant ($F_{1,14} = 0.07$, $P = 0.80$), and the `HatchOrder` main effect is significant ($F_{1,14} = 13.04$, $P = 0.0028$), the statistical significance of `Diet` is once again marginal ($F_{1,14} = 4.07$, $P = 0.063$). After controlling for hatching order, the predicted sex ratios from this model are 0.61 on the restricted diet and 0.45 on the *ad libitum* diet, suggesting that rationing food leads to a 35% increase in the proportion of males in the brood. Thus, even though the effect of `Diet` was statistically nonsignificant, the magnitude of the apparent effect suggests that it would be premature to discount the effect of diet on sex ratio in zebra finches.

3.5.3.2 Conclusions

As with all GLMs, logistic regression allows the simultaneous testing of several interacting factors and covariates in a single model. Since the underlying error distribution of sex ratios is known (or presumed) to be binomial, this can be explicitly incorporated into the modelling process, so avoiding *ad hoc* transformations and nonparametric tests which lack power. By weighting sex ratios by their denominators, each individual contributing to the ratio is given equal significance. In contrast to Kilner's (1998) analyses, we found no statistical evidence that diet was a significant determinant of sex ratio in zebra finches (though nonsignificance was marginal). This conclusion was independent of whether hatching order was included or omitted as a covariate in the model. However, in accord with Kilner, we found that females tend to hatch before males. This result was independent of adult feeding regime and, because it utilized information from all of the chicks that hatched successfully, the robustness of our conclusion is illustrated by the high significance of the result ($P < 0.001$).

3.5.4 A case history: Opossum sex ratios

Austad and Sunquist (1986) carried out the first manipulative field test of the Trivers–Willard prediction (Chapter 13) that mothers in relatively good condition will produce more male-biased sex ratios than poor-condition mothers. Females of the common opossum, *Didelphis marsupialis* (a polygynous marsupial producing litters of 2–12 offspring), were given either diet supplements or no supplements (control) and the sexes of subsequent offspring were recorded. Austad and Sunquist analysed these data by comparing the overall sex ratio produced by females in the treatment group with that produced by the control group using a one-tailed binomial test. They found a significant difference ($P = 0.007$)

in sex ratio (but not in litter size) between the two groups.

Subsequently, Wright *et al.* (1995) correctly pointed out that comparing the overall sex ratios of the two groups was inappropriate since the hypothesis under test predicts an individual level response, not a population (treatment group) level response, to maternal condition. They reanalysed the data (as presented in Sunquist & Eisenberg 1993) using litters as the sampling unit. Litters were classified categorically as 'male-biased' or 'unbiased or female-biased'. Using a χ^2 test they found no significant difference between the proportions of male-biased litters produced by control and supplemented females (18/36 and 20/36 respectively; $\chi^2 = 0.068$, $P > 0.05$).

However, Wright *et al.*'s categorization of litter sex ratio does not use all of the available information since the actual composition of each litter (the size of the litter and the degree of any sex ratio bias) is overlooked. For example, litters containing six males plus one female are placed in the same 'male-biased' category as litters of four males plus three females, while the degree of male bias is different (see also Box 3.1). Similarly, litters of three males plus one female and litters of six males plus two females have the same sex ratio and are treated equally, despite the fact that larger litters give more trustworthy sex ratio estimates (section 3.3.3.4).

In an attempt to arrive at a more robust conclusion, the opossum data (as obtained from ME Sunquist) were explored using weighted logistic analyses (Hardy 1997). In a first analysis, litters were the sampling unit (*D. marsupialis* produces two cohorts of litters per season) and litters produced by the same mother were assumed to be statistically independent. No significant influence of cohort was found, so litters produced by the same mother were lumped and a second analysis was performed with mothers as the sampling unit: intuitively, this is more appropriate since the assumption of independent litter sex ratios does not have to be made, and it was mothers, not litters, that received the experimental treatments. Due to overdispersion, Williams' adjustment (appropriate when the binomial denominator varies, section 3.4.4.2) was employed and significance was evaluated with one-tailed *t*-tests. Both analyses found that the sex ratio produced by food-supplemented mothers was significantly more male-biased than the sex ratios of control females' offspring (e.g. second analysis, $t = 1.973$, $df = 40$, $P = 0.028$; mean sex ratio of supplemented group = 0.577, control group = 0.488).

One-tailed tests were used because there was an anticipated direction for any difference between treatment groups (i.e. H_0: 'there is no sex ratio difference between the two groups'; H_1: 'the offspring sex ratios of supplemented females are more male-biased than those of control females'). However, not all statisticians agree that one-tailed tests can be used when deviations in the unanticipated direction are possible (Rice & Gaines 1994). Using two-tailed tests would have led to the acceptance of H_0, but would have been suspiciously close to significance at the 5% level (second analysis, $P = 0.0566$). See Hardy (1997) for further discussion, including the use of 'directed tests' (Rice & Gaines 1994) as an alternative intermediate to the extremes of one- and two-tailed testing.

Sven Krackow (pers. comm.) recently reanalysed the opossum data using both GLMs and generalized linear mixed modelling (GLMMs, section 3.4.4.2) which includes a random between-litter effect. For these data, the analysis leads to the same biological conclusion regardless of whether a GLMM or a corrected GLM is employed (Krackow opted for two-tailed testing and concluded lack of significance, $P > 0.061$ for both analyses), while employing uncorrected GLMs led to spurious significance ($P < 0.04$).

Regardless of the degree of statistical significance, the effect of diet supplements on sex ratio is not exceptionally large ('supplemented' litters contained 18% more males), suggesting that more data are probably required before we can reach satisfactorily firm conclusions. Problems in researching mammalian sex ratios are discussed in Chapter 13.

3.6 | Simulation studies

We have argued that GLMs (and their 'offspring') are usually the most appropriate analyses for sex ratio analyses. In this section, we challenge this argument using a series of simulations to ask

two questions. First, under what circumstances are errors likely to be generated when analysing sex ratio data? Second, when does the method of analysis really matter?

We use simulated datasets generated using the *rbinom* procedure in *S-Plus* and compare three methods of analysis:

1. A nonparametric method (Wilcoxon Rank Sum Test, equivalent to Mann–Whitney *U*-test).
2. A transformation method (*t*-test on arcsine-squareroot transformed data).
3. A generalized linear model with binomial errors.

In each case, sex ratios are expressed as proportions. (For a comparable analysis for negative binomial data, see Wilson *et al.* 1996 and Wilson & Grenfell 1997.)

3.6.1 Simulation approach

Imagine we want to determine whether females in a population of burying beetles exhibit a clutch sex ratio response to some manipulated variable (e.g. change in day length). We could randomly assign the beetles to one of two treatment groups (increasing day length or decreasing day length) and record the sex ratios of the broods produced (see also section 3.5.4). What is the probability that we will incorrectly *reject* the null hypothesis of no difference between the treatment groups (type I error) or incorrectly *accept* the null hypothesis (type II errors)?

To address this question, we randomly generated two datasets, representing the two treatment groups. In each case, we produced a series of 'virtual' clutches of a given size and sex ratio drawn from the binomial distribution. We then used our three methods to test for a significant difference between the two groups, and repeated this process 1000 times. Thus, the probabilities of type I and type II errors are, respectively, equal to the proportion of simulations in which the analysis indicated a significant difference between the two groups when there wasn't one, or no significant difference when there was one. We performed these simulations for a number of different scenarios. For example, to examine the effect of clutch size on the probability of making errors, we allowed clutch size to vary between 1 and 20 eggs per female, and to determine the effect of sample size we varied the number of broods included in the analysis between 10 and 50 per treatment group. Sex ratios were allowed to vary between 0 and 1.

3.6.2 Effect of clutch size, sample size and sex ratio

In these simulations, the probability of making a mistake was qualitatively similar for all three methods (K Wilson unpublished analyses). Therefore, in Figure 3.6 we show the results just for the GLM model. This figure comprises 12 graphs, representing the results of the combined effects of clutch size (1, 5, 10 and 20) and sample size ($n = 10$, 20 and 50). Each graph is divided into an 11 × 11 matrix and the axes of the matrix represent the mean sex ratio of each of the two treatment groups (varying between 0.0 and 1.0 in intervals of 0.1). Each cell of the matrix is colour-coded depending on the probability of making an error; the darker the cell, the higher the probability of making a mistake. Thus, white cells represent instances in which there is 0–10% average probability of making a mistake and deep-red cells indicate that the probability of making a mistake is 90–100%. Type I errors are indicated by the colour of cells on the leading diagonal of each matrix (bottom-left to top-right), and type II errors are indicated by the colour of the remaining cells.

Examining just the leading diagonals of each matrix (bottom-left to top-right), it is fairly clear that the probability of making a type I error (i.e. detecting a spurious difference between treatments) remains at less than 10%, regardless of clutch size, sample size or mean sex ratio. The biggest effects are seen in the probability of making a type II error (i.e. failing to detect significant differences between treatments). As expected, the probability of making a type II error is reduced when clutch sizes are large, sample sizes are large and the effect of our manipulation on sex ratio is large. In other words, we make fewer mistakes when the power of our test is high!

3.6.3 Differences between statistical methods

What about quantitative differences between the three methods? Simulations indicated that the nonparametric test and the *t*-test on arcsine-transformed data differ relatively little in their

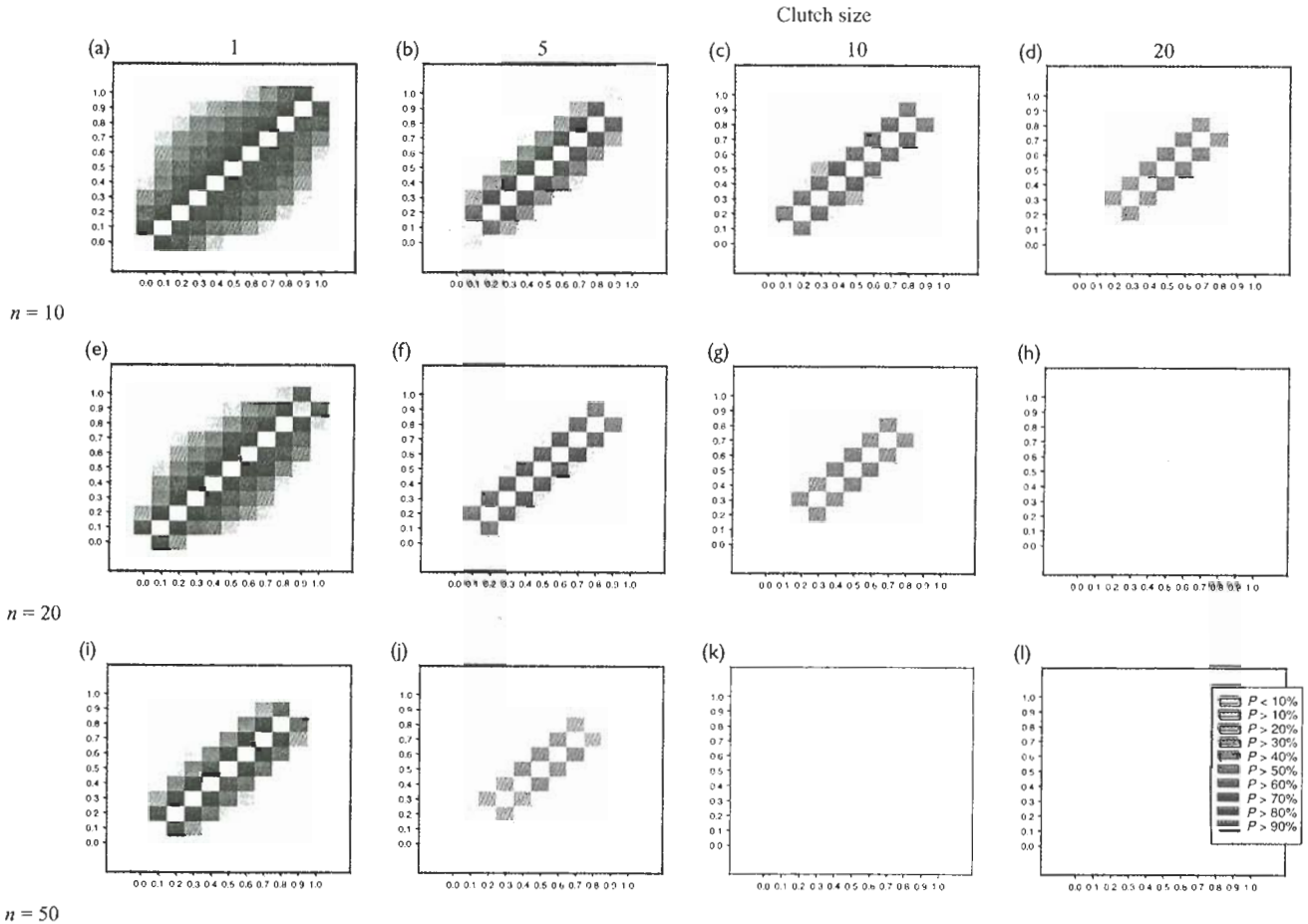

Fig 3.6 Probability of type I and type II errors as a function of difference in sex ratio, clutch size and sample size. Each plot (a – l) represents the results of a series of 1000 simulations examining the ability of a statistical model to discriminate between two datasets comprising 'virtual' clutches of a given size and sex ratio drawn from the binomial distribution. The vertical and horizontal axes on each plot are the mean sex ratios of each of the two clutches being compared (varying between 0 and 1 in steps of 0.1). Clutch size was allowed to equal 1, 5, 10 or 20; sample size equalled 10 clutches, 20 clutches or 50 clutches. The probability of making a mistake was qualitatively similar for all three methods (see text) and so here we show the results just for the GLM model. Each cell of the matrix is colour-coded depending on the probability of making an error; the darker the cell, the higher the probability of making a mistake (see legend attached to plot l). Type I errors are indicated by the intensity of colour in the cells of the leading diagonal of each matrix (bottom-left to top-right), and type II errors are indicated by the intensity of colour in the remaining cells.

higher the probability of making a mistake (see legend attached to plot I). Type I errors are indicated by the intensity of colour in the cells of the leading diagonal of each matrix (bottom-left to top-right), and type II errors are indicated by the intensity of colour in the remaining cells.

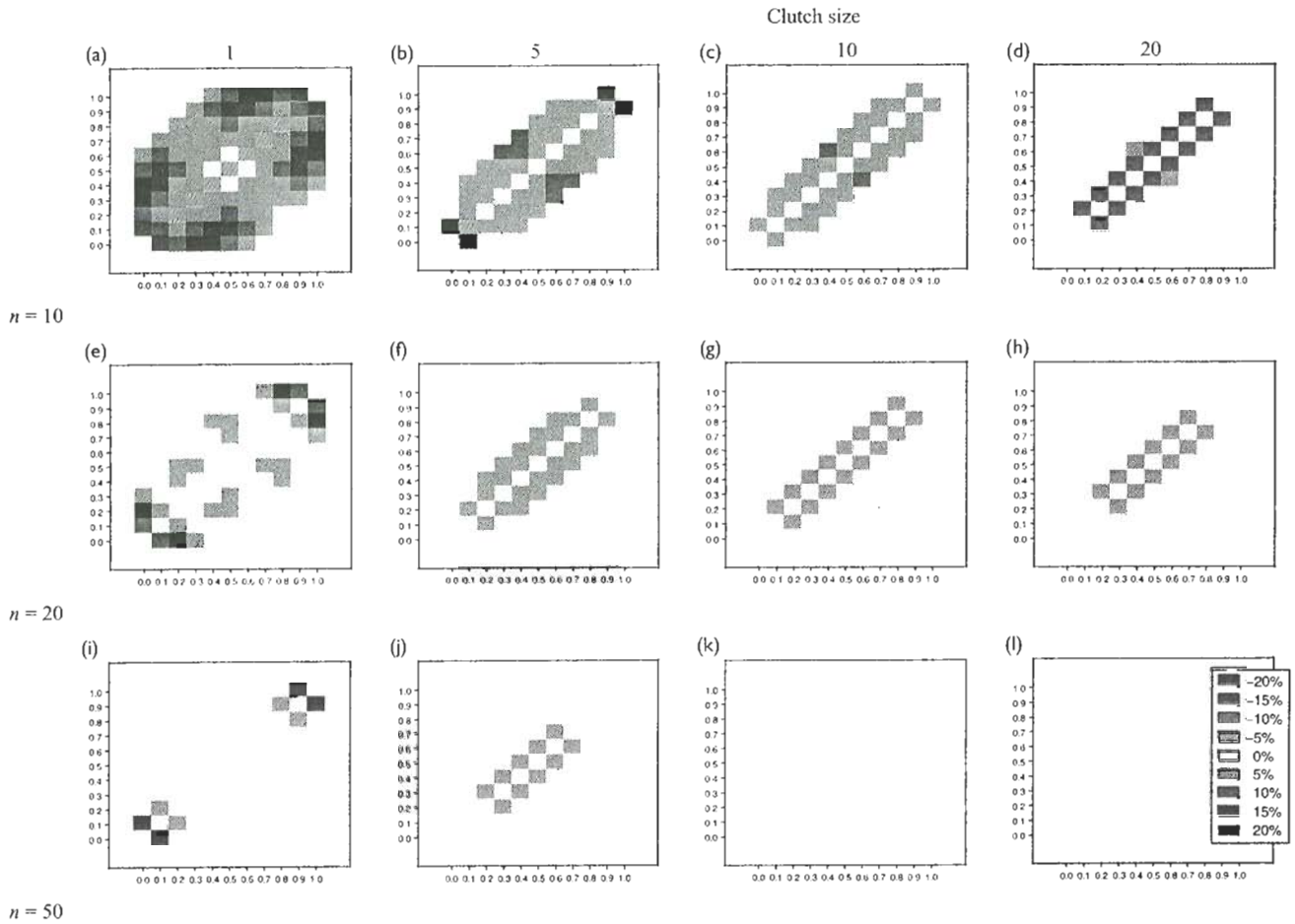

Fig 3.7 The difference between two statistical methods for analysing sex ratio data in their probability of generating type I and type II errors. For details of the simulations, see the legend to Figure 3.6 and the text. In this figure, quantitative differences between the nonparametric and GLM models are shown. Here, the different colours represent differences between the nonparametric method and the GLM in their probability of producing errors (see the legend attached to plot I). When there is little difference between the two methods (i.e. 0–5% difference in the number of errors) the cell is white; when the nonparametric method is better (i.e. produces fewer errors) the cell is coloured pink or red, and when the GLM is better the cell is coloured various shades of blue. In general, there is little difference between the two methods in their probability of producing type I errors (the leading diagonals tend to be white, except when clutch size is one and sample size is very small, i.e. 10), but the GLM method produces significantly fewer type II errors, except when sample sizes are large (outside the leading diagonal, there is much more blue than red).

probability of generating errors (K Wilson unpublished analyses), whereas these two methods differ quite markedly from the GLM with binomial errors. This point is illustrated in Figure 3.7. Here, the different colours represent differences between the nonparametric method and the GLM in their probability of producing type I and type II errors. When there is little difference between the two methods (i.e. less than 5% difference in the number of errors) the cell is coloured white; when the nonparametric method is better (i.e. produces fewer errors) the cell is coloured pink or red, and when the GLM is better the cell is coloured various shades of blue. This figure suggests that there is very little difference between the two methods in their probability of generating type I errors (cells in the leading diagonal are generally coloured white). However, the two methods differ greatly in their probability of generating type II errors: genuine differences between treatments are much less likely to be detected using the nonparametric method than when using the GLM (i.e. as expected, the nonparametric method lacks power-efficiency; section 3.3.1.1). It is also apparent that the benefit of using the GLM approach is generally enhanced when clutch and sample sizes are small. Interestingly, it appears that when clutch sizes are small ($\leq$5) the difference between the two methods is greatest when both mean sex ratios are close to 0 or 1, whereas when clutch sizes are large ($\geq$10) the benefits of using the GLM approach are most evident when both sex ratios are close to 0.5.

Thus, these simulations indicate that sex ratio differences are likely to be difficult to detect in species with small clutch sizes, except when sample sizes are large. Moreover, although type I errors appear to be unlikely in sex ratio analyses regardless of the analytical method used, type II errors are much less likely when using logistic regression than when using alternative methods, especially when clutch and sample sizes are small.

3.7 | Conclusions

The most appropriate approach for analysing sex ratios (and other proportion data) will often be logistic regression (GLM with binomial errors and logit-link function). After all, sex ratios are expressed as proportions and logistic GLMs were developed to analyse proportion data. We hope to have shown that using GLMs is not very much (if at all) more complex than using classical parametric methods (which are currently the most frequently used). The next time you have collected a set of sex ratio data and are ready to begin analysis, ask yourself whether you want to make best use of the data. If you do, your initial approach should be to use GLMs.

Acknowledgements

We thank Tim Benton, Mick Crawley, Simon Gates, Fiona Hunter, Sven Krackow, Kate Lessells, Evert Meelis, Gösta Nachman, Gordon Smyth and Ian Stevenson for help, discussion and comments; any mistakes or misunderstandings are ours not theirs! We thank Darren Shaw for the use of some of his *S-Plus* functions and Becky Kilner and Fiona Hunter (and co-workers) for providing us with some of their raw data. We thank both them and Mel Sunquist, Pat Weatherhead and their co-workers, for tolerating the intrusion of re-analysis. This chapter was written whilst K Wilson was funded by the Natural Environment Research Council, UK. ICW Hardy was funded by the European Commission and the University of Sunderland.

A3.1 | Reference sources and statistical packages for GLMs

The most accessible book on GLMs (i.e. a book written by a biologist rather than a statistician) is probably Crawley's (1993) *GLIM for Ecologists*, which has recently been superseded by Crawley's (2002) *Statistical Computing*. More detailed statistical background can be found in books by Aitkin *et al.* (1989), Cox and Snell (1989), Hosmer and Lemeshow (1989), McCullagh and Nelder (1989), Agresti (1990), Dobson (1990), Collett (1991) and Menard (1995). In addition to these general reference sources, biologist-friendly descriptions of logistic analysis are provided by Shanubhogue and Gore (1987), Trexler and Travis (1993), Sokal and

Rohlf (1995) and Hardy and Field (1998). Some recent examples where GLMs have been used to analyse complex sex ratio datasets are discussed in Hartley *et al.* (1999). Wilson *et al.* (1996) and Wilson and Grenfell (1997) give accounts of GLMs with particular reference to analysing parasite count data.

Logistic analysis is available in at least the following packages: *BIOM-pc*, *BMDP*, *EGRET*, *Genstat*, *GLIM*, *GLIMStat*, *JMP*, *LOGXACT*, *MacAnova*, *SAS*, *S-Plus*, *SPSS*, *SPSSX*, *STATA*, *STATISTIX* and *SYSTAT*. Agresti (1990) provides an appendix detailing the options available in various packages and advice on their implementation. The manuals for some of these packages are also excellent reference resources (e.g. SAS Institute Inc. 1995, SPSS 1999). While we de not recommend the *GLIM* manual (Francis *et al.* 1993) for the nonprofessional, the *GLIM* package itself becomes much more user-friendly if you have a copy of Crawley (1993). The following website provides updated information about the most frequently used statistical packages:
http://www.maths.uq.edu.au/~gks/webguide/statcomp.html.

References

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

Aitkin M, Anderson D, Francis B & Hinde J (1989) *Statistical Modelling in GLIM*. Oxford: Clarendon Press.

Austad SN & Sunquist ME (1986) Sex-ratio manipulation in the common opossum. *Nature*, **324**, 58–60.

Cohen J (1988) *Statistical Power Analysis for the Behavioural Sciences*, 2nd edn. London: Lawrence Erlbaum Associates.

Collett D (1991) *Modelling Binary Data*. New York: Chapman & Hall.

Cox DR & Snell EJ (1989) *Analysis of Binary Data*, 2nd edn. New York: Chapman & Hall.

Crawley MJ (1993) *GLIM for Ecologists*. Oxford: Blackwell Scientific Publishing.

Crawley MJ (2002) *Statistical Computing*. London: John Wiley.

Dobson AJ (1990) *An Introduction to Generalized Linear Models*, New York: Chapman & Hall.

Elston DA (1998) Estimation of denominator degrees of freedom of *F* distributions for assessing Wald statistics for fixed effect factors in unbalanced mixed models. *Biometrics*, **54**, 1085–1096.

Finney DJ (1971) *Probit Analysis*. Cambridge: Cambridge University Press.

Fisher RA (1954) *Statistical Methods for Research Workers*, 12th edn. Edinburgh: Oliver & Boyd.

Flanagan KE, West SA & Godfray HCJ (1998) Local mate competition, variable fecundity and information use in a parasitoid. *Animal Behaviour*, **56**, 191–198.

Francis B, Green M & Payne C (eds) (1993) *The GLIM System: Release 4 Manual*. Oxford: Oxford University Press.

Hardy ICW (1997) Opossum sex ratios revisited: significant or nonsignificant? *American Naturalist*, **150**, 420–424.

Hardy ICW & Field SA (1998) Logistic analysis of animal contests. *Animal Behaviour*, **56**, 787–792.

Hardy ICW & Mayhew PJ (1998) Sex ratio, sexual dimorphism and mating structure in bethylid wasps. *Behavioral Ecology and Sociobiology*, **42**, 383–395.

Hartley IR, Griffith SC, Wilson K, Shepherd M & Burke T (1999) Nestling sex ratios in the polygynously breeding Corn Bunting *Miliaria calandra*. *Journal of Avian Biology* **30**, 7–14.

Hosmer DW & Lemeshow S (1989) *Applied Logistic Regression*. New York: Wiley.

Hunter FM, Jones IL, Wilson K, Dawson DA & Burke TA (in prep.) Manipulation of offspring sex in species with mutual sexual selection.

Kilner R (1998) Primary and secondary sex ratio manipulation by zebra finches. *Animal Behaviour*, **56**, 155–164.

Krackow S & Tkadlec E (2001) Analysis of brood sex ratios: implications of offspring clustering. *Behavioral Ecology and Sociobiology*, **50**, 293–301.

Kruuk LEB, Clutton-Brock TH, Albon SD, Pemberton JM & Guinness FE (1999) Population density affects sex ratio variation in red deer. *Nature* **399**, 459–461.

Leonard ML & Weatherhead PJ (1996) Dominance rank and offspring sex ratios in domestic fowl. *Animal Behaviour*, **51**, 725–731.

Leonard ML & Weatherhead PJ (1998) Erratum. *Animal Behaviour*, **55**, 777.

Lipsey MW (1990) *Design Sensitivity: Statistical Power for Experimental Research*. London: Sage Publications.

Mathsoft Inc (1999) *S-Plus Guide to Statistics*, volume 1. Seattle: Mathsoft. Inc.

McCallum H (2000) *Population Parameters: Estimation for Ecological Models*. Oxford: Blackwell Science.

McCullagh P & Nelder JA (1989) *Generalized Linear Models*, 2nd edn. New York: Chapman & Hall.

Meddis R (1984) *Statistics Using Ranks: A Unified Approach*. Oxford: Blackwell.

Menard S (1995) *Applied Logistic Regression Analysis*. Sage University paper series on quantitative applications in the social sciences, 07–106. Thousand Oaks: Sage.

Neave HR & Worthington PL (1988) *Distribution-free Tests*. New York: Routledge.

Petersen G & Hardy ICW (1996) The importance of being larger: parasitoid intruder-owner contests and their implications for clutch size. *Animal Behaviour*, **51**, 1363–1373.

Rice WR (1990) A consensus combined *P*-value test and the family-wide significance of component tests. *Biometrics*, **46**, 303–308.

Rice WR & Gaines SD (1994) 'Heads I win, tails you use': testing directional alternative hypotheses in ecological and evolutionary research. *Trends in Ecology and Evolution*, **9**, 235–237.

Rohlf FJ & Sokal RR (1995) *Statistical Tables*, 3rd edn. New York: WH Freeman & Co.

SAS Institute Inc (1995) *Logistic Regression Examples Using the SAS System*, Version 6, 1st edn. Cary, NC: SAS.

Shanubhogue A & Gore P (1987) Using logistic regression in ecology. *Current Science*, **20**, 933–935.

Siegel S & Castellan NJ (1988) *Nonparametric Statistics for the Behavioural Sciences*, 2nd edn. New York: McGraw-Hill.

Sokal RR & Rohlf FJ (1995) *Biometry*, 3rd edn. New York: WH Freeman & Co.

SPSS (1999) *SPSS Base 9.0 User's Guide*. Chicago: SPSS Inc.

Sunquist ME & Eisenberg JF (1993) Reproductive strategies of female *Didelphus*. *Bulletin of the Florida Museum of Natural History, Biological Sciences*, **36**, 109–140.

Trexler JC & Travis J (1993) Nontraditional regression analyses. *Ecology*, **74**, 1629–1637.

Westerdahl H, Bensch S, Hansson B, Hasselquist D & VonSchantz T (1997) Sex ratio variation among broods of great reed warblers *Acrocephalus arundinaceus*. *Molecular Ecology*, **6**, 543–548.

Williams DA (1982) Extra-binomial variation in logistic linear models. *Applied Statistics*, **31**, 144–148.

Wilson K & Grenfell BT (1997) Generalized linear modelling for parasitologists. *Parasitology Today*, **13**, 33–38.

Wilson K, Grenfell BT & Shaw DJ (1996) Analysis of aggregated parasite distributions: a comparison of methods. *Functional Ecology*, **10**, 592–601.

Wright DD, Ryser JJ & Kiltie RA (1995) First-cohort advantage hypothesis: a new twist on facultative sex ratio adjustment. *American Naturalist*, **145**, 133–145.

Zar JH (1999) *Biostatistical Analysis*, 4th edn. New Jersey: Prentice Hall Inc.