3

Model Parameter Estimation

3.1 Models and Data

3.1.1 Fitting Data to a Model

A mathematical model of a biological population is always a simulation of nature. When a model is descriptive or explanatory of a situation in nature, it is necessary to estimate values for at least some of its parameters by optimizing the fit between the expectations from the model and data observed from nature. Parameter estimation is fundamental to the science of modelling populations, and this chapter focuses on the different ways that data can be fitted to a model.

The design of a model, such that it adequately represents the structure of the modelled system, primarily relates to determining which variables are to be included (i.e., whether the model will include age structure, relate to numbers or biomass, etc.) and the relationships between them (linear, nonlinear, etc.). Model specification is clearly a vital step in modelling any system, and one tends to proceed by development from a simple model to a more complex representation. Development of model complexity tends to stop due to a lack of data rather than a desire to keep a model simple.

Once the model has been designed, if it is not just a simulation model, then it remains to be fitted, or have its parameters "tuned" to the observable world, using whatever data are available from the system. Fitting a model to data implies estimating values for the model's parameters to optimize the agreement between the model's predictions and the data to be fitted from nature.

When fitting a model to observed data, three things are required:

1. A formal mathematical model of the system, which is capable of generating predictions about the observable world
2. Data from the system to be used when estimating the parameter values
3. A criterion (sometimes called a merit or objective function) to judge the quality of fit between the model's predictions and the observed data

An example could be a fishery stock assessment model that predicts changes in catch rates and age structure through time. Fitting the mathematical model to nature entails varying trial values of the model parameters until an optimum agreement is found, according to the criterion selected, between the model predictions of how catch rates and catch-at-age will change through time and the observed time series of real information. However, as described in Chapter 1, there is more to the process of modelling than the three formal requirements listed earlier. If this were not the case, then one would invariably finish with a multiparameter empirical model providing the best statistical fit, but not being interpretable in a realistic way. If an explanatory model is wanted, then we need a rather obvious but non-quantifiable extension to the third requirement of the quality of fit criterion. As well as optimizing some best fit criterion, the optimum model should also give rise to biologically sensible predictions. It can happen that a model fit that generates an optimal mathematical solution still predicts biological nonsense (e.g., predictions of an enormous initial population size and almost no productivity, with the history of fishing being one of the gradual erosion of the accumulated biomass). Not all such deviations from biological reality are so obvious, so we need to guard against a lack of realism in the outcomes of fitting a model. In short, we generally want our models to be simple but realistic. It comes down to how one selects which model to use and what is meant by quality of fit.

3.1.2 Which Comes First, the Data or the Model?

We should always be asking the question: Given a particular model, with particular parameter values, how well does it predict (how likely are) the observed data?

The process of model fitting has two parts. First, a set of one or more models is selected or designed, and second, values for the model's parameters are found that optimize the quality of fit according to the formal criterion selected by the modeller. Except for at least some of the variables used, model selection is independent of the data (the variables we observed, against which the model is to be compared, must obviously be included in the model). Thus, we always determine how well the selected model(s) can emulate the available data. Once a particular model has been selected, the optimum parameter values for that model are determined by the fixed set of observed data. In fact, what tends to happen is that one starts with a relatively simple representation of whatever is being modelled, and that is fitted to available data, extending and articulating the model in steps.

The process of positing a model plus parameter values and comparing its implications against nature reflects the hypothetico-deductive approach to scientific practice (Popper, 1963; Lakatos, 1970). However, remember, from Chapter 1, that fitting a model to data only tests how successfully it can describe the data, not how well it explains the data.

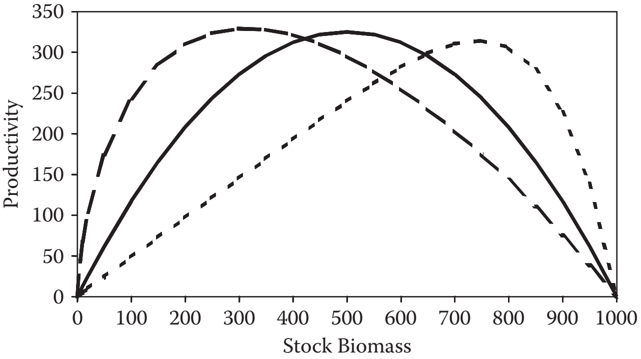

Fig ur e 3.1
Three different biomass production curves (cf. Chapter 2). The equation used was productivity = $(r/p)*B((1 - (B/K)^p)$, where r is a growth rate, K is the equilibrium biomass, B is the stock biomass, and p is the coefficient of asymmetry. When $p = 1$ (solid line), the equation simplifies to the standard logistic Schaeffer production curve, symmetric about $K/2$. When p is less than 1, the production curve has a mode lower than $K/2$ ($p = -0.25$, dashed line), and when p is greater than 1, the mode is to the right of $K/2$ ($p = 7$, dotted line). Whenever $p < 1$ or > 1, the curve is asymmetric and needs to be rescaled to have a absolute maximum productivity similar to that of the Schaeffer curve.

3.1.3 Quality of Fit versus Parsimony versus r eality

With a set of structurally different models, there are criteria other than just quality of numerical fit that should be used to determine the preferred model. For example, consider the difference between two models of stock production; the first assumes a symmetric production curve against stock size (linear density dependence) while the second has no such restriction (Figure 3.1). The asymmetric production curve is an improvement over the symmetric in terms of both biological reality and mathematical generality (because it is less restricted in its biological interpretation and because it contains the symmetric model as a special case when $p = 1$).

Using the asymmetric model enhances the potential for realism and probably the quality of fit, but there has also been an increase in model complexity. If the quality of fit between two models were equivalent (i.e., $p \sim 1.0$), one would tend to use the most realistic or the simplest. The general question to be answered is: Do the benefits derived from adding extra parameters to a model outweigh the losses and thereby justify their addition? There is no simple answer to this question because the realism of a model is not quantifiable.

Adding parameters to a realistic model may convert it into a multiparameter empirical model (cf. Figure 1.6). Extra parameters are likely to improve the quality of fit of most models because they permit greater flexibility in the model outcomes. In the extreme, one could have the same number of parameters as one had data points and could obtain a perfect fit of the model to the data. Such a model would be both *ad hoc* and useless, but it would certainly fit the data points well. Clearly, the quality of fit between a model and data is not everything.

Selecting an optimum explanatory model requires a trade-off between improving the quality of fit between the model and the data, keeping the model as simple as possible, and having the model reflect reality as closely as possible. Increasing the number of parameters will generally improve the quality of fit but will increase the complexity and may decrease the reality. The latter quality is the hardest to assess.

Ignoring the problem of whether a model is realistic, quantitative measures have been developed that assess the balance between the relative quality of fit (the variation in the data accounted for) and the number of parameters fitted. Following our intuitions, these measures suggest that parameter addition be rejected if the improvement in quality of fit is only minor. Such measures include likelihood ratio tests and Akaike's information criterion (AIC; see Burnham and Anderson, 1998); these will be introduced after we have considered maximum likelihood methods.

A further problem with selecting an optimum fit relates to data quality. If the available data are limited in quantity or quality (sadly this is common when modelling fisheries), the number of parameters that can be estimated adequately is also limited. The data might be said to be uninformative about those extra parameters. Adding extra parameters to a model that can already be fitted to data reasonably well may make the fitting procedure unstable. A more complex model that provides a significantly better fit may be possible, but the solution may be less precise or more biased. Punt (1990) appeared to have this problem with data on Cape hake (Haddon, 1998).

3.1.4 u ncertainty

Irrespective of how closely one can fit a model to a set of data, there is no logical or automatic implication that the model chosen is necessarily the best representation of the system being modelled. In addition, there is no guarantee that the model will accurately predict how the system will behave in the future, or even that the parameter estimates are the optimum values for the system. Very often, it is possible to obtain essentially equally good fits to one's data with sometimes widely different sets of parameters (roughly, the more parameters being estimated, the more likely this is to occur). This is especially the case when there are strong correlations between parameters. The correlations act to offset the effects of changing one parameter, as altering the values of others means that different sets of parameter values describe the available data almost equally well. The different sets may even produce similar predictions. Model selection is especially difficult under such circumstances. When the results of a model fitting exercise are this uncertain, it suggests that the data are insufficiently informative to distinguish between the alternative possibilities. Uncertainty is an unpleasant commonplace in stock assessment, and how best to approach it is a vital part of fisheries modelling (Francis, 1992; Punt and Hilborn, 1997; see Chapter 8).

It is undoubtedly valuable knowing how to estimate model parameters, but the value of this is greatly increased if we also have some way of determining the confidence with which we can use the estimated parameter values and other model outputs. We need to be able to characterize the uncertainty inherent in our analyses. The uncertainty in any analysis stems both from the data, which will contain random variation from observation errors and variation due to random factors, and from model uncertainty, where the model fails to capture the processes and stochasticity in nature. How to characterize this uncertainty will be addressed in Chapter 8.

3.1.5 Alternative Criteria of g oodness of Fit

Three criteria of model fit are commonly used today. Most biologists new to modelling will only be familiar with the least squared residual error approach. We will consider methodologies that use least squares, maximum likelihood, and Bayesian statistics as their criterion of model fit. The choice of the criterion of fit remains a controversial field (Dennis, 1996; and associated papers). Some researchers stoutly defend a maximum likelihood approach, but others imply that if an assessment is not conducted using a Bayesian approach, it must be next to useless. It is hoped that this chapter will illustrate that much of this controversy is misplaced, and that the choice of criterion of fit should depend upon the objectives of the modelling process. The assessment strategy to use should be whichever is most convenient and can do the job required without distorting the results.

3.2 Least Squared Residuals

3.2.1 introduction

The most commonly used criterion of fit is still the one known as *least squares*. It is so called because it involves a search for parameter values that minimize the sum of the squared differences between the observed data (e.g., time series of catch rates or catch-at-age) and the predictions from the combination of model and particular parameter values. Typical fisheries data would never exactly fit a proposed model, even if that model were correct. The differences between the models' predictions and data observations are known as residual errors (sometimes called noise on the signal). A statement made about a model's error structure is making a claim about the expected statistical distribution of the residual error around each of the predicted observations. It is important to understand that when using the least squared

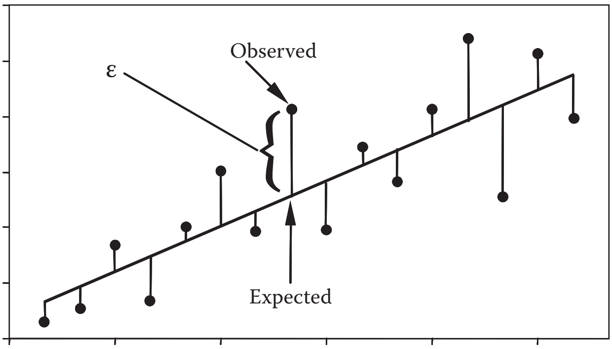

Fig ur e 3.2
View of the residual errors for a linear relation between two variables. The error terms represented are those for the regression of *Y* on *X*, where the observed and expected values are on the *Y* axis ($Y = a + bX + \varepsilon$) for given values of the *X* variable. The residuals for a regression of *X* on *Y* would be horizontal. The residual errors (the εs) are the differences between the expected values of the *Y* variable, given by the regression line, and the observed values. Clearly, for a well-fitted line they will be both positive and negative in value.

residual criterion of model fit, the residual errors are always assumed to be normally distributed. The term $N(\mu,\sigma^2)$ is the standard nomenclature for such normal, random residual errors, and should be read as implying a normal distribution with a mean of μ and a variance of σ^2. In general terms, observed *value* = predicted *value* + ε, or observed – predicted = ε, where ε (epsilon) is the residual or random error. Strictly, the phrase used should always be "residual error," but it is very common to use just one or the other term (*residual* or *error*) interchangeably. Note that the residual error term is added and not multiplied to the predicted (fitted or expected) value (Figure 3.2). It is common to use the two terms (*residual* and *error*) to describe one concept and three terms (*predicted*, *fitted*, and *expected*) to describe another single concept. By pointing out the equivalence within these sets of terms, it is hoped that possible confusion might be avoided.

Normal residuals can be either positive or negative (Figure 3.2). Thus, it would not suit our purpose just to search for parameter values that would minimize the sum of the residuals, as the negative values would cancel some of the positive in an *ad hoc* manner. However, it would be a real option to minimize the sum of the absolute values of the residuals, although this is not particularly convenient mathematically (Birkes and Dodge, 1993; see Example Box 3.1).

Squaring each residual error removes the problem of negative residuals and is mathematically very convenient. With absolute residuals each value would have equal weight, whereas squaring each residual gives extra weight to larger residuals and less weight to residuals less than one. For this reason, absolute residuals are sometimes used to implement model fitting methods that are robust to outliers (Sokal and Rohlf, 1995). Such differences in

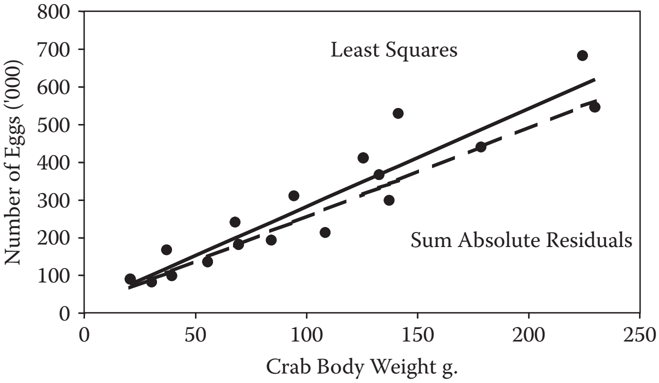

Fig ur e 3.3
Number of eggs vs. body mass for the New Zealand swimming crab *Ovalipes catharus*. (Data from Haddon, 1994.) The upper line is the best fitting least squares linear regression (eggs = 25.51 + 2.599*Wt), while the lower dashed line is the best fitting least sum of absolute residuals (eggs = 17.08 + 2.37*Wt). The lines differ because the error structures used to fit the lines are different and the lines cannot be directly compared statistically. Both are optimal fits according to their respective criteria (see Example Box 3.1). Most people tend to prefer the line that keeps the data around it in a symmetrical fashion.

emphasis are why one obtains a different "optimal" fit, depending on which criterion is used (Figure 3.3, Example Box 3.1).

The objective when using the least squares criterion is to minimize the sum of the residual errors squared. From this we gain:

$$SSQ = \sum \left(Observed - Expected \right)^2 \tag{3.1}$$

where SSQ is the sum of the squared residuals, e.g., for a linear regression, $SSQ = \Sigma[Y - (a + bX)]^2$. It should be remembered that there is an analytical solution for the simple least squares linear regression, which can be found in any statistical text (e.g., Snedecor and Cochran, 1989; Sokal and Rohlf, 1995). It is invariably more efficient to use an analytical solution to a problem, and one should do so wherever it is possible.

A major assumption of least squares methodology is that the residual error terms exhibit a normal distribution about the predicted variable with equal variance for all values of the observed variable; that is, the σ^2 in the $N(0,\sigma^2)$ is a constant. If data are transformed in any way, the transformation effects on the residuals may violate this assumption. Conversely, a transformation may standardize the residual variances if they vary in a systematic way. As always, a consideration or visualization of the form of the residuals is good practice.

The sum of absolute residuals (SAR) is intuitively attractive in that it gives all data points equal weight. However, at least in the linear model, it is a fact

EXAMPLE BOX 3.1

Fitting a straight line using the sum of absolute residuals (SAR) and the sum of squared residuals (SSQ). Size relates to drained body weight in grams and eggs is the fecundity in thousands for the crab *Ovalipes catharus* (data selected from Haddon, 1994). PredSSQ and PredSAR are the predicted values of eggs using the parameters defined for the two different criteria of fit, respectively. Totals represents the respective sums of both types of residual errors. The intercepts and gradients are clearly different (Figure 3.3). The solutions are found in each case using the solver, minimizing E3 and F3 in turn. The solution for the SAR criterion is highly unstable; try different starting points and compare the apparently optimal solution found. Compare this with the relative stability of the least squares criterion. The SSQ solution must pass through the mean of each variable. The SAR solution must pass through two of the points. Plot columns A to D as in Figure 3.3. Which line fit looks best to your eyes? Body size and fecundity for rows 11 to 21 are (69.46, 181.713), (84.12, 193.161), (94.31, 310.425), (108.47, 213.247), (125.54, 411.056), (132.7, 366.567), (137.31, 298.439), (141.34, 529.351), (178.60, 440.394), (224.31, 683.008), and (229.89, 545.681).

	A	B	C	D	E	F
2		**Intercept**	21.50974	17.07990	**Totals**	
3		**Gradient**	2.59998	2.37019	62764.18	=sum(F5:F21)
4	**Size**	**Eggs**	**PredSSQ**	**PredSAR**	**SSQ**	**SAR**
5	20.71	89.35	=C$2+A5*C$3	=D$2+A5*D$3	=(B5-C5)^2	=abs(B5-D5)
6	30.35	82.399	=C$2+A6*C$3	=D$2+A6*D$3	=(B6-C6)^2	=abs(B6-D6)
7	37.04	166.97	=C$2+A7*C$3	=D$2+A7*D$3	=(B7-C7)^2	=abs(B7-D7)
8	39.50	98.324	*Copy*	*down*	*to*	*Row 21*
9	55.60	135.427	166.07	148.86	938.92	13.44
10	67.90	240.713	198.05	178.02	1820.25	62.70

that the best fitting line must always pass literally through two of the points (Birkes and Dodge, 1993; also see Figure 3.3). This has the disadvantage of sometimes forcing the residuals to be asymmetric.

It should be noted that the standard least squares linear regression is a regression of Y on X. This means that the X values are assumed to be measurable with no errors and to be independent of the Y values, while the Y values are assumed to be dependent upon the given X values. In the fitting process, this implies the residuals would be vertical (Figure 3.2). Of course, in most biological processes it would not be possible to measure the so-called independent variable without error. In an ideal world, it would perhaps be best to have residual errors that were perpendicular to the expected line (i.e., neither Y on X or X on Y), as in functional regression or principal components analysis. Ricker (1973) and McArdle (1990) discuss this problem at length. Generally, the bias that using Y on X might introduce is likely to be small, although it would become greater as the variability about the predicted curve becomes greater.

3.2.2 Selection of r esidual error Structure

A common alternative criterion of quality of fit is *maximum likelihood*. Parameters are selected that maximize the probability density or likelihood that the observed values (the data) would have occurred given the particular model and the set of parameters selected.

One great advantage of the maximum likelihood approach is that it forces one to be explicit about the statistical form of the expected residual errors, that is, whether they are normal and additive, as with the regression examples we have seen ($Y = a + bX + \varepsilon$), or lognormal and multiplicative (as in $Y = aX^{b}e^{\varepsilon}$), or follow some other distribution (more on this later). Whichever error structure is selected, with least squares it is necessary to devise a linearizing transformation to convert the selected error structure into normally distributed residuals. Thus, with lognormal errors, a log transformation will permit linear, least square methods (e.g., $y = ax^{b}e^{\varepsilon}$ becomes $\mathrm{Ln}(y) = \mathrm{Ln}(a) + b\mathrm{Ln}(x) + \varepsilon$). If no normalizing transformation is possible, or if it fails to linearize the modelled relation, then it becomes necessary to use nonlinear methods to fit the models to data. In such cases, if one still used normal residuals, this would greatly influence the optimum fit. If the residuals required are not normal and there is no normalizing transformation, then the least squares approach is not an option, but one could use maximum likelihood methods. Generally, it must be remembered that with least squares, the selection of the residual error structure is implicit, and at worst, it is *ad hoc*.

We will consider maximum likelihood methods in detail after we have introduced nonlinear parameter estimation methods and expanded on the method of least squares.

3.3 Nonlinear Estimation

3.3.1 Parameter estimation Techniques

Our examples have so far been limited to the estimation of parameters for simple linear models. Using least squares, it would be simpler to use the analytical solution for linear regression rather than invoking the solver facility built into Excel. Of course, most fisheries models are far more complex, involving many more parameters and nonlinear relationships between the variables. There are usually no simple analytical solutions or linearizing transformations available for these more difficult to fit models. Numerical methods are needed when fitting these multiparameter, nonlinear models to data. In order to understand the strengths and weaknesses of such numerical methods, it is necessary to have some understanding of the algorithms or strategies used in the more commonly available methods used for fitting complex models. When no analytical solution exists for fitting a model to data, we need to search for the optimum parameter values. We can define three types of search: graphical searches, directed searches, and heuristic searches. To understand nonlinear parameter estimation, we will first consider the graphical approach to searching for optimal parameter values.

3.3.2 g raphical Searches for Optimal Parameter Values

Consider the simple problem of fitting a curve to the relationship between carapace width and eggs carried per batch in the crab *Ovalipes catharus* (Figure 3.4; Haddon, 1994). The values exhibited by the data points suggest the exponential relationship

$$\text{Eggs} = a.e^{b.\text{CWidth}}e^{\varepsilon} \tag{3.2}$$

where Eggs refers to fecundity, CWidth refers to the carapace width, and *a* and *b* are the parameters to be estimated. Because the residual errors in Equation 3.2 are lognormal, it can be linearized by natural logarithmic transformation:

$$\text{Ln}\left(\text{Eggs}\right) = \text{Ln}\left(a\right) + b.\text{CWidth} + \varepsilon \tag{3.3}$$

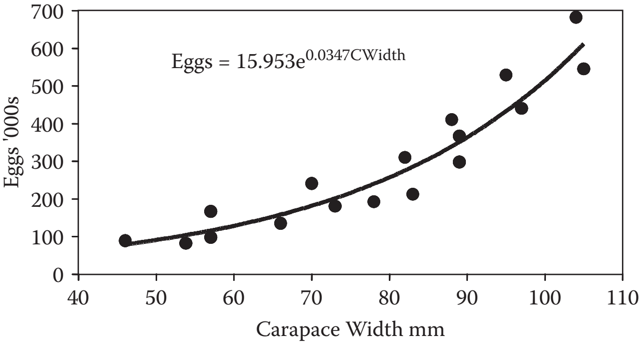

Fig ur e 3.4
The exponential relationship between number of eggs in an egg mass and carapace width (CWidth) for the New Zealand crab *Ovalipes catharus*. (Data are a subset from Haddon, 1994.) The exponential relationship describes over 90% of the variation in the data. The solid line illustrates the given optimal solution (see Example Box 3.2).

EXAMPLE BOX 3.2

Fitting an exponential curve to eggs at carapace width data. Width is in mm and eggs is fecundity in thousands for *Ovalipes catharus* (cf. Figure 3.4; data are from Haddon, 1994). The Predicted column is the expected number of eggs. SSQ represents the sum of squared residuals (in C1). The solution is found using the solver, minimizing C1 by changing C2 and C3; an exact solution can be found using the array function Linest, but that would require a separate column of transformed egg numbers (try this; don't forget to use <Ctrl><Shift><Enter> to enter the array function). Note the log transformation of the predicted number of eggs in column D (see Equation 3.3). By plotting column B against A as separate points, and then exp(column C) against column A as a connected line, you should be able to mimic Figure 3.4. Use this sheet to investigate the relative precision of different solutions. The values in columns E and F represent different trial values of Ln(a) and b with their respective sum of squared residuals. Paste the values of your trial values and SSQ into spare cells in columns E and F; i.e., copy B1:C3 to keep a record of your trials. Do different parameters that give a similar SSQ alter the graph visually? How precise should one try to be when using biological data of limited original accuracy? For rows 10 to 21 the data for width and fecundity are (70, 240.713), (73, 181.713), (78, 193.161), (82, 310.425), (83, 213.247), (88, 411.056), (89, 366.567), (89, 298.439), (95, 529.351), (97, 440.394), (104, 683.008), and (105, 545.681).

	A	B	C	D	E	F
1		**SSQ**	**=sum(D5:D21)**		**Trial Values**	
2		**Ln(a)**	**2.7696**		**1)** SSQ	0.630620
3		**b**	**0.034734**		**1)** Ln(a)	2.7696
4	**Width**	**Eggs**	**Predicted**	**ResidSQ**	**1)** b	0.034734
5	46	89.35	=C$2+A5*C$3	=(Ln(B5)-C5)^2		
6	53.8	82.399	=C$2+A6*C$3	=(Ln(B6)-C6)^2	**2)** SSQ	0.630620
7	57	166.97	=C$2+A7*C$3	=(Ln(B7)-C7)^2	**2)** Ln(a)	2.77
8	57	98.324	*Copy down*	*To row 21*	**2)** b	0.03473
9	66	135.427	5.062	0.024		

Equation 3.3 has the form of a linear regression, which can be solved analytically, leading to Ln(a) = 2.7696 and b = 0.0347. By back-transforming the Ln(a), i.e., Exp(2.7696), we obtain a = 15.953, which produces a satisfactory fitted line (Figure 3.4; see Example Box 3.2).

An alternative approach would be to carry out a grid search across plausible parameter values and plotting the resulting sum of squared residuals as a contour plot (Figure 3.5). This defines a valley or pit in the sum of squares

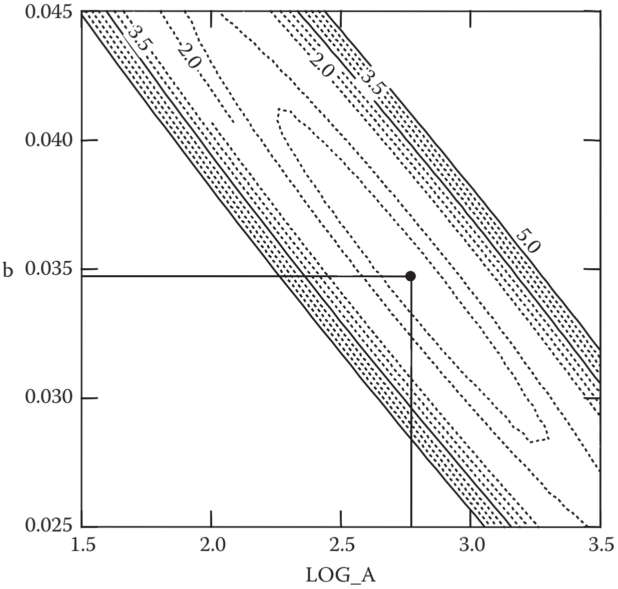

Fig ur e 3.5
Graphical search for the optimal parameter values for Equation 3.5 given the data in Example Box 3.2. Different combinations of the two parameters produce different values of the sum of squared residuals. Plotting these as contours makes it possible to home in on the optimum values. The optimal solution is indicated (at 0.0347 and 2.769).

surface of possible values and the optimal parameter values relate to the bottom of the pit. Obviously, the graphical search grid must bracket the minimum sum of squares.

The contour plot makes it possible to constrain the parameter values being used in the search for the minimum least squares. This is a downhill search for the bottom or minimum of the surface. Each of the contours represents combinations of parameter values that give rise to the same value of sum of squared residuals, meaning they provide fits of equal quality.

Instead of a simple trial-and-error search for the minimum, it is possible to use this analogy of a downhill search and, using information gained from previous trials, conduct an informed trial-and-error search. The contour map indicates visually how to improve the parameter estimation. This entails moving the trial parameters to form combinations that are most likely to lead to a maximal reduction in the sum of squares (the bottom of the pit). This informed trial and error is continued until there is no detectible improvement to the sum of squares.

As we have seen from the contours in Figure 3.5, the same sum of squares can be obtained from different combinations of parameters. If we contemplate the sixth significant digit in the parameter estimates, then the optimum fit would be graphed as a point on a very small contour circle. If there is no analytical solution and the model has to be fitted using some numerical method, then the optimal fit is a compromise between the time taken to find the solution and the accuracy of the fit. In Example Box 3.2, the impact of different trial values of the two parameters on the predicted values and graph (Figure 3.4) can be determined.

The graphical search is informative in a number of ways. The shape of the contour plot provides information about relationships between the parameters of the model. If the parameters were completely independent of one another (the ideal but rare case), then the contours would form either perfect circles or vertical ovals. If the contours were ovals, but at some angle, then parameter correlation is indicated (e.g., Figure 3.5).

3.3.3 Parameter Correlation and Confounding effects

The resolution of the contours shown in Figure 3.5 is sufficient to indicate approximately where the "best" fit solution would lie. The fact that the sum of squared residual values forms a diagonal trough (left to right) indicates that there is a negative correlation between the two parameters. This means we cannot readily distinguish between different combinations of the two parameters. What it implies is that if we were to force a particular value of a on the model, we could still recover a reasonable fit because the value of b could be altered accordingly to produce a similar curve. The plot provides us with a visual indication of the quality of fit and the confidence we can have that our best fit provides a good indication of the actual situation.

Very strong parameter correlation is a problem that occurs with many fisheries statistics. For example, the total mortality for a fish population is a combination of natural and fishing mortalities, and there is always difficulty in separating these different phenomena. Parameter correlations imply that the so-called independent variables (X axis variables) cannot be independently determined and, in fact, are often strongly correlated. When supposedly independent variables are correlated, it becomes impossible to determine which "independent" variable is most closely related to changes in the so-called dependent variable; their effects are said to be confounded.

Sadly, the graphical search strategy is of limited use for models with more than three parameters. In the example of the fecundity vs. carapace width there were only two parameters, and the sum of squared residuals can be visualized as a third axis described by the a and b *p*arameters to form a third dimension to generate a surface. With a three-parameter model the criterion of fit values would be described as a four-dimensional volume or three surfaces. With n parameters the criterion of fit would be described by an $n + 1$ dimensional hypervolume. Even with a hypervolume the analogy of moving downhill or over the surface of the criterion of fit is still a useful intuition to hold in mind.

The basic idea and strategy of the downhill search lies behind many of the numerical methods for fitting nonlinear models to observed data, even with *n* parameters. With the least squares criterion, the mathematical problem is one of minimization of the sum of the squared residuals. To move beyond the graphical search, some other, more efficient numerical method is required. The graphical approach is essentially a grid search that could be pursued with more and more precision once the minimum values had been bracketed. Automatic minimization routines do something similar without the need for the grid or human intervention. There are numerous algorithms, but here we will only describe the directed and heuristic searches.

3.3.4 Automated Directed Searches

A common approach, known as the Levenberg-Marquardt algorithm, uses the downhill analogy directly (Press et al., 1989). It requires a single set of starting parameter values to begin the search. These guessed parameters specify a particular point on the sum of squared residuals (SSQ) surface (which has the same number of dimensions as there are parameters, and so may be a hypervolume).

Despite the number of dimensions, the analogy remains of a surface having a steepest path downwards to a minimum. By automatically making tiny increments to each of the parameter values in turn, the algorithm can estimate the gradient of the surface along each of the parameter dimensions (a form of numerical partial differentiation). The algorithm then alters the parameters in the direction that should lead to the steepest decline down the SSQ surface. Given this new point on the surface, the relative gradients are considered once again from the new viewpoint and the parameters incremented again in whichever direction will lead to the greatest decline in the SSQ. This is continued until a minimum is reached or further benefits are minimal (Press et al., 1989). The algorithm automatically directs the new trial values along a path that should maximize the decline in the criterion of fit.

An obvious problem with this search strategy is the possibility of local minima on the SSQ surface that could be confused for the global minima by a non-linear-fitting algorithm (Figure 3.6). An equally obvious test of the generality of any solution is to start the search with initial parameter guesses that widely bracket the final estimated values. They should all converge on the same answer if it is to be considered an adequate estimate. Another potential problem, which is becoming less important, is that a large model with many parameters may take a very long time to converge on a possible solution for all parameters (Schnute et al., 1998). More likely is that the nonlinear solver will fail through instability of some form.

The solver in later versions of Excel is surprisingly robust and is useful for many smaller problems. For larger problems, it may be necessary to resort to custom computer programs or to use a meta-programming language

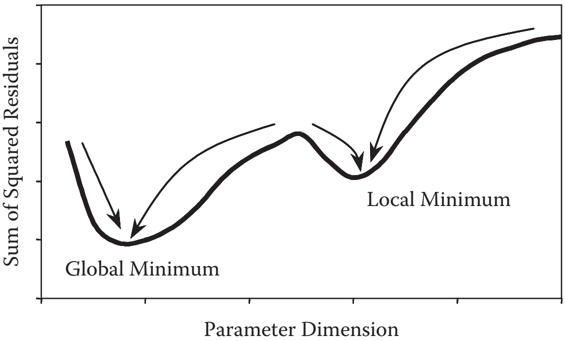

Fig ur e 3.6
Schematic representation of a single-parameter dimension with a local minimum that could confuse an automatic minimization routine. Once in the vicinity of the local minimum, every change in the parameter value makes the SSQ value get larger, and many routines would understandably stop and claim to have found a minimum.

such as AD-Model Builder (Schnute et al., 1998; see www.admb-project.org; AD-Model Builder is now open source).

3.3.5 Automated Heuristic Searches

A robust alternative to the Levenberg-Marquardt algorithm is the Simplex algorithm. This is an $n + 1$ dimensional bracketing search method (Nelder and Mead, 1965) where there are n parameters to be estimated. This method requires $n + 1$ sets of initial trial values that attempt to bracket the optimal solution or at least define the direction in which to move the set of trial values over the surface of the criterion of fit. By comparing the relative fit of the $n + 1$ different sets of trial parameter values, the direction that should lead to maximum improvement in fit can be determined approximately. $n + 1$ sets of trial values are required so it becomes possible for the overall set to bracket the optimum combination. The $n + 1$ set of trial values move over/through the n-dimensional hypervolume, always in the approximate direction that should improve the value of the criterion of fit. Some speak of the $n + 1$ dimensional search object "crawling" over the fitting criterion surface toward the minimum. The analogy with an amoeba flowing through n-dimensional space has been used (Press et al., 1989). This continues until the only way the set of trial values can improve the fit is for the $n + 1$ parameter combinations to contract toward each other (i.e., they really bracket the optimum solution and contract toward it). The simplex algorithm is very robust but can be slow, depending on the complexity of the model and how close the starting values are to the optimum.

Whichever approach is used, one should get to know the limitations of any particular nonlinear solver that might be used, and always attempt to find the same solution from a number of different sets of initial parameter guesses.

3.4 Likelihood

3.4.1 Maximum Likelihood Criterion of Fit

Maximum likelihood, used as the criterion of quality of fit, is usually characterized as the determination or search for the set of model parameters that maximize the probability that the observed values (the data) would have occurred given the particular model and the set of parameters selected (Press et al., 1989; Neter et al., 1996). Using maximum likelihood requires the model to be defined so that it specifies probabilities or likelihoods for each of the observations (the available data) as a function of the parameter values and other variables in the model. To obtain the likelihoods, one needs to define how the residual errors are distributed about the expected values derived from the model. To understand the idea of maximum likelihood parameter estimation we need first to formalize the idea of using probability distributions. Many different statistical probability distributions can be used to describe residual error structures. We will consider the normal, lognormal, binomial, Poisson, gamma, and multinomial distributions as well as some of their uses in fisheries modelling. There are many other distributions that could be used (Elliott, 1971; Hastings and Peacock, 1975).

3.4.2 The Normal Distribution

Most biological scientists would understand a claim that a set of observations is expected to exhibit a normal distribution about their expected value (the mean). What this implies is that the observed values of a variable X are expected to be distributed symmetrically about their mean, and that large deviations from the mean would occur less often than small deviations. The expected relative rates of occurrence of different-sized deviations from the mean are described by the probability density function (pdf) for the normal distribution (Equation 3.4, Figure 3.7). Relative frequency histograms count the occurrence of individuals in a population that are found in defined classes of the variable under study (e.g., body weight). A pdf for the normal distribution describes, in effect, the expected relative frequency (probability density) curve generated for a continuous variate instead of for discrete classes of the variate. The pdf for the normal curve has two parameters, the mean or expectation of the distribution (μ) and its standard deviation (σ). Once they are set for a variable X, substituting different values of the variable into the equation generates the well-known normal curve (Figure 3.7, Equation 3.4, Example Box 3.3).

$$\text{Probability Density} = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{-(X-\mu)^2}{2\sigma^2}\right)} = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right)} \tag{3.4}$$

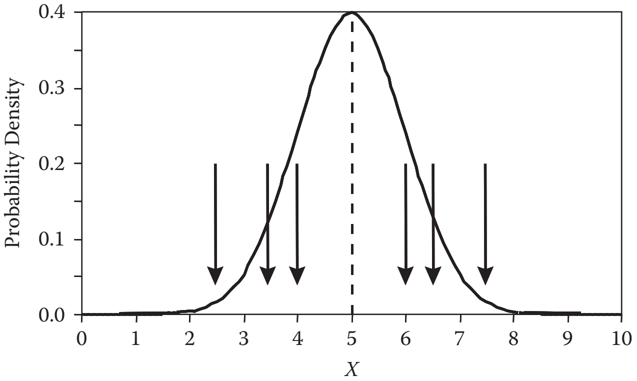

Fig ur e 3.7
The arrows represent a set of six observations, 2.5, 3.5, 4.0, 6.0, 6.5, and 7.5, with a mean of 5.0 and standard deviation of approximately 1.95. The probability density function (pdf) of a normal distribution with a mean of 5.0 and standard deviation of 1.0 (Equation 3.4), from which they are hypothesized to have been sampled, is superimposed about them. Note the symmetry. From Equation 3.4, with a mean of 5.0 and standard deviation of 1.0, the pdf value of *X* at 2.5 and 7.5 is 0.018, at 3.5 and 6.5 is 0.130, and at 4 and 6 is 0.242. These values are the probability densities of the respective *X* values given the selected parameter values. See Example Box 3.3.

For any given value of the variable *X*, the value of this function (Equation 3.4) defines its probability density. To someone unfamiliar with it, this equation may look rather daunting, but it can be implemented easily in most spreadsheet programs (see Example Box 3.3) or on a hand calculator. The right-hand version of the pdf (Equation 3.4) demonstrates that each observation is being converted into a residual and is then standardized by dividing by the standard deviation ($[(X - \mu)/\sigma]$, i.e., observation *X* minus the mean, μ, divided by the standard deviation). In other words, with the normal distribution, we are determining the probability density of particular values considered in terms of how many standard deviations these values are from the mean (look ahead to Figure 3.9).

3.4.3 Probability Density

Note the use of the phrase "probability density" instead of probability. The sum of the probabilities of the full set of possible outcomes for a particular event must equal 1. Thus, when tossing an unbiased coin, the possible outcomes are a head or a tail, each with a probability of 0.5, and these add to 1 for any particular coin-tossing event (a discrete variable). With a continuous variable, an event would be the making of a single observation *X*. However, speaking mathematically, and ignoring such things as the limits of measurement, with a truly continuous variable there are an infinity of potential observed values of *X* within the range of what is possible. The sum of the probabilities of all possible events (i.e., *X* between $-\infty$ to $+\infty$) must equal 1, and

EXAMPLE BOX 3.3

Properties of the normal probability density function. In cell E1 put the equation =1/(B2*sqrt(2*PI())); this is the constant. In cell B5 (and copy down to row 105, where X = +5 in column A), put Equation 3.4: =$E$1*exp(–((A5-$B$1)^2)/(2*$B$2^2)). Alternatively, you could use the Excel function, normdist(A5,$B$1,$B$2,false), which provides the same answer without having to bother with the constant in E1. Check the help for a description of this function. In C5 (and copy down), put the function normdist(A5,$B$1,$B$2,true), to obtain the cumulative probability density. The last number in column C should be very close to 1. The sum in B3 should be close to 10, which reflects the increments of 0.1 that were used to step through from –5 to 5. If you reconstruct the worksheet to increment the values in column A by 0.05 instead of 0.1, don't forget to modify cell B3. What happens to the value in B3 when you make this change? Standardize the separate probability densities by dividing each one by their sum, as in column D. Cumulate column D in column E; it sums to 1. Column C doesn't quite reach 1 because there is a small but finite probability of values being greater than 5, given a mean of 0 and standard deviation of 1. Plot columns B and C against column A to see the standard normal curve and the cumulative normal curve (cf. Figure 3.9). Alter the values in B1 and B2 and see the impact on the curves. See the text for the difference between probability density (column B) and probability. Find the probability densities for the six X values in Figure 3.7.

	A	B	C	D	E
1	**Mean**	0		**$1/\sigma\sqrt{2\pi}$**	0.39894228
2	**StDev**	1			
3	**Sum**	=Sum(B5:B105)			
4	**X**	**Prob. Density**	**Cumulative Prob. Density**	**Standardized**	**Cumulative**
5	–5	1.4867E–06	2.871E–07	=B5/$B$3	=D5
6	–4.9	2.4390E–06	4.799E–07	=B6/$B$3	=D6+E5
7	–4.8	3.9613E–06	7.944E–07	3.9613E–07	7.887E–07
8	–4.7	6.3698E–06	1.302E–06	6.3698E–07	1.426E–06
9	.	*Copy down*	*Copy down*	*Copy down*	*Copy down*
105	5.0	1.4867E-06	0.9999997	1.4867E-07	1

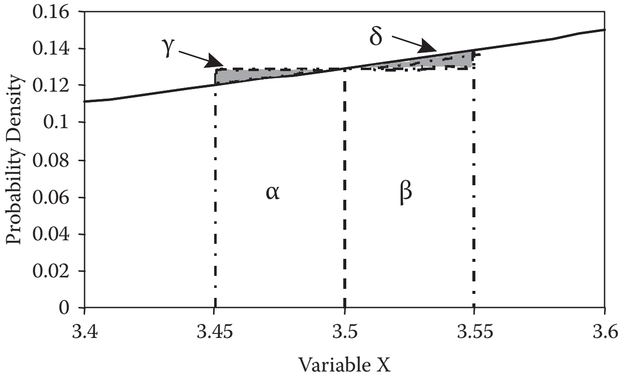

Fig ur e 3.8
Probability densities for a normally distributed variate *X* having a mean of 5.0 and a standard deviation of 1.0 (as in Figure 3.7). The pdf value at $X = 3.5$ is 0.129518. The true area under the curve between 3.45 and 3.55 is $\alpha + \beta + \delta$; an approximation to this is $\alpha + \beta + \gamma$ (see Example Box 3.4).

so, with an infinity of possibilities, the probability of any particular value (e.g., exactly 2.5) would be infinitesimally small. Instead, to obtain literal probabilities for continuous variates, we need to quantify an area under the pdf. Thus, we can have a probability for a range of a continuous variate *X* but not for a particular value.

Clearly there is a difference between probability density and probability because, given a known mean and standard deviation (Figure 3.7), the probability of observing an *X* value of exactly 3.5 is infinitesimal, but using the given parameter values, the probability density of exactly 3.5 is 0.130. To grasp the ideas behind using likelihoods, it is necessary to understand this difference between probability density and probability (Example Boxes 3.3 and 3.4).

The graph of the pdf values (Figure 3.7) is analogous to a histogram of expected relative proportions for a continuous variate. This reflects our intuitions with regard to the relative chance of obtaining different values of the continuous variable *X*. Thus, with our example (Figure 3.7) and the given parameter values, the pdf value for an observation of 4 (pdf = 0.242) is thirteen times greater than (more likely than) the pdf value for an observation of 2.5 (pdf = 0.018). We can use this characterization to determine whether the observed data are consistent with the hypothesized normal pdf curve from which they are assumed to be sampled.

The term *density* is used as an analogy to express the weight or mass of probability above particular values of the variable *X*. Of course, when the range over which each density operates is infinitesimal, the overall probability would also be infinitesimal. However, consider what would be the case if,

EXAMPLE BOX 3.4

Estimating probabilities under the normal distribution. The calculations involved reflect those seen in Figure 3.8. P(3.45–3.55) refers to the probability of the variable lying between 3.45 and 3.55 and equals $\alpha + \beta + \delta$. The terms α and $(\beta + \gamma)$ are derived from the normdist function. The approximate probability can be derived from the single-value probability density for 3.5 multiplied by the area concerned, and the difference between the strict probability and the approximate (i.e., cells B6 and D6) can be seen. Note we need to use six decimal places to detect the difference. The use of the probability density in the approximation only becomes valid once we multiply it by the range over which it is intended to apply and thereby generate an area under the normal curve.

	A	B	C	D
1	**3.45**	=normdist(A1,5,1,true)		
2	**3.5**	=normdist(A2,5,1,true)	=normdist(A2,5,1,false)	
3	**3.55**	=normdist(A3,5,1,true)	$\alpha + \gamma = \beta$	**=C2*0.05**
4	α	=B2–B1		
5	$\beta+\delta$	=B3–B2		
6	**P(3.45–3.55)**	=B4+B5	**Approx. P(3.45–3.55)**	=2*D3
7	δ	=B5–D3	$\alpha + \beta + \delta =$ **P(3.45–3.55)**	=B4+D3+B7
8	γ	=D3–B4		

as an approximation, one were to assume the same probability density operated over a small range of the variable of interest (Figure 3.8). Using the approximation that the probability density at $X = 3.5$ represents the probability density at values of X close to 3.5, the probability (as contrasted with probability density) of observing a value between 3.45 and 3.55 is the area $\alpha + \beta + \gamma = 0.129518 \times 0.1 = 0.0129518$. This is only an approximation to the real area under the pdf curve between those ranges (which is $\alpha + \beta + \delta = 0.0129585$). The difference, or error, is $\delta - \gamma$, and this would become smaller and smaller as the range over which the approximation was applied became smaller. The limit would be where the range was infinitesimal, whereupon the solution would be exact but also infinitesimal (Example Box 3.4).

Instead of summing under the curve in a series of discrete, small steps (which provides only an approximation, as in Figure 3.8 and Example Box 3.4), it is better to sum the area under the pdf curve in an infinitesimal way using integration. The assertion that the sum of the probabilities of the set of all possible outcomes for any given observation cumulates to 1 can be expressed by integrating Equation 3.4 between the maximum and minimum possible values for the variable X (Equation 3.5):

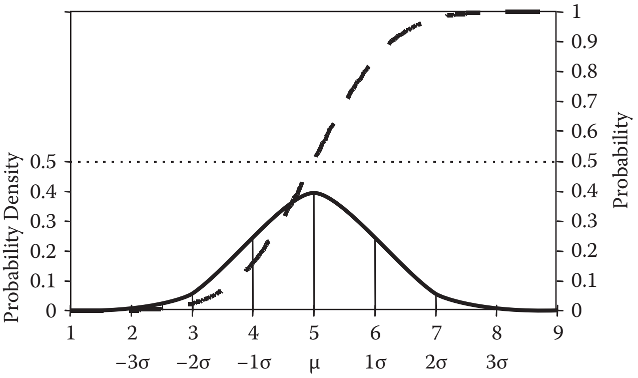

Fig ur e 3.9
The relationship between the probability density function (solid line) and its integral (or the cumulative distribution function) for the normal distribution (in this instance with a mean of 5.0 and a standard deviation of 1.0). Integration is equivalent to summing the area under the pdf so we see that it sums to 1 across all possibilities. We can also see that the symmetry of the normal pdf leads to an observation being the mean or less than the mean, having a probability of 0.5.

$$\text{cdf} = 1 = \int_{X=-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{-(X-\mu)^2}{2\sigma^2}\right)} \tag{3.5}$$

where cdf refers to the cumulative distribution function. This provides an exact solution to the area under the pdf and translates the probability densities into true probabilities (Figure 3.9).

Integration of the pdf can be performed across more ranges of the variate to produce the probability of observing values within a given range (Example Box 3.4). For example, if we integrated the pdf between *X* values of 5 and 4, we would be calculating the probability of making an observation with a value between 4 and 5 (Figure 3.10). To do this does not require an exercise in calculus each time, as there are published tables of the cumulative normal frequency distribution (e.g., Table A3 in Snedecor and Cochran, 1967). Using such tables or the equivalent functions in Excel, one can determine the cdf value for 5.0 and subtract from it the cdf value for 4.0 to produce the area under the pdf between the limits of 4 and 5 (Figure 3.10).

The values we obtain by substituting real values into Equation 3.4 are known as probability densities, and these are infinitesimally summed or integrated between given limits to produce probabilities, as in Equation 3.5. Thus, in terms of true probability, with respect to continuous variates, pdf values can only be interpreted directly when referred to in the context of a range of the possible values (e.g., Figure 3.10). Later, when we consider discrete variates (e.g., the binomial distribution—heads or tails, tagged or

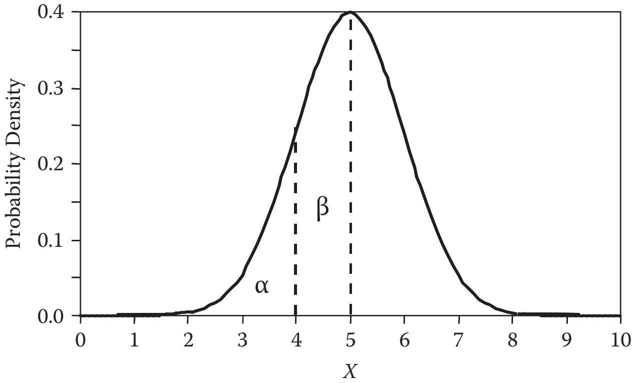

Fig ur e 3.10
A normal distribution pdf with a mean of 5 and standard deviation of 1. The total area under this curve from -∞ to +∞ sums to 1. As the curve is symmetrical, we can conclude that the probability of observing a value of less than or equal to 5 would be 0.5. Therefore, the area α + β = 0.5. To obtain the probability of a value falling between 4 and 5, we need the probability of obtaining a value of 4 or less (i.e., =normdist(4,5,1,TRUE); α = 0.158655), and this must be subtracted from the cdf for the value of 5 or less to leave the probability in which we are interested, i.e., β (= 0.341345).

untagged), the pdfs generate discrete probabilities. With discrete variate pdfs, probability density and probability are the same because the number of possible outcomes to any particular event is limited by definition.

3.4.4 Likelihood Definition

With continuous variates the true probability of individual observations is infinitesimal. However, we have also repeated a statement found in many statistical texts about likelihood being the probability of the observed data given a set of parameters. This apparent anomaly requires clarification. With any model of a process there will be parameter sets for which the model predictions are obviously inconsistent with the available data and are thus unlikely. Conversely, there will be other parameter sets that produce predictions closely consistent with the data, and these we feel are far more likely. Maximum likelihood methods are attempting to find the parameter set that is most likely in this sense of the word *likely*. Press et al. (1989) present a clear statement of the problem of applying likelihood methods to parameter estimation.

> It is not meaningful to ask the question, 'What is the probability that a particular set of fitted parameters $a_1 \ldots a_M$ is correct?' The reason is that there is no statistical universe of models from which the parameters are drawn. There is just one model, the correct one, and a statistical universe of data sets that are drawn from it! (Press et al., 1989, p. 549)

What they are claiming is that a set of observations only constitutes a sample of what is possible. If we were able to take another sample, we would expect to obtain similar, but not exactly the same, values. This is because our sampling is based upon the assumption that there is some underlying explanatory model for the behaviour of the system, and this constrains the measurable variate's behaviour to follow the predicted model output plus the residual terms defined by the pdf of the errors. If this is the case, then, as stated, there is only one correct model and only one set of correct parameters. This means that each different set of parameter values is a separate hypothesis and not a sample from some statistical distribution of parameter values. Press et al. go further and state:

> We identify the probability of the data given the parameters … as the *likelihood* of the parameters given the data. This identification is entirely based upon intuition. It has no formal mathematical basis in and of itself. (Press et al., 1989, p. 550)

This appears to be a very weak foundation upon which to base the serious business of model parameter estimation. However, one will often see either a definition or an implied definition of likelihood that is very exact:

$$L(\theta) = \prod_{i=1}^{n} pdf\left(X_i \middle| \theta\right) \tag{3.6}$$

which is read the likelihood the parameter(s) θ (theta) is the product of the pdf values for each of the *n* observations X_{i} given the parameter(s) θ. It is common knowledge that when events are independent, it is necessary to multiply their separate probabilities to obtain the overall probability (e.g., the probability of three heads in a row is 0.5 × 0.5 × 0.5 = 0.125). With continuous variates, the same process is involved, so we use the product (capital Pi, ∏) of the separate probability densities and not the sum. The product of all the separate pdf values when the parameters are set equal to the hypothesized values is called the likelihood value and is usually designated *L*(θ) (Neter et al., 1996). Thus, in our earlier example (Figure 3.7), we could assume a standard deviation of 1 and then search for the value of the mean that would lead to the maximum likelihood. We have observations at 2.5, 3.5, 4.0, 6.0, 6.5, and 7.5. If we were to assume a range of different values for the mean, we would find that the respective pdf values for each observation would alter so that their product, the likelihood for the particular guessed value of the mean, would also vary. Clearly, to search for the most likely value of the mean, we need to try different values and search for that value which maximizes the product of the pdf values (Example Box 3.5, Figure 3.11).

EXAMPLE BOX 3.5

Maximum likelihood (ML) search for an estimate of the mean value of a set of observations. Columns B to D are example trials (Figure 3.11). Row 9 contains the product of each set of likelihoods. Using the solver, maximize the value in E9 by changing cells E1:E2. The StDev of the data, =StDev(A3:A8) = 1.94936, is larger than the estimate in E2. The ML estimate of the standard deviation is =sqrt(Σ(x-μ)^2/n), the divisor is n and not n–1. In a separate cell calculate the usual StDev and beside it put =sqrt((E2*E2*6)/5). Are they the same? This difference between the usual population estimate of the standard deviation and the ML estimation of the same statistic becomes important when we want to use ML to fit data to a model and need to estimate the unbiased variance.

	A	B	C	D	E
1	Mean	4.5	5	5.25	5
2	Obs\StDev	1	1	1	1.7795
3	2.5	0.05399	0.01753	0.00909	=normdist($A3,E$1,E$2,false)
4	3.5	0.24191	0.12952	0.08628	=normdist($A4,E$1,E$2,false)
5	4	0.35207	0.24197	0.18265	=normdist($A5,E$1,E$2,false)
6	6	0.12952	0.24197	0.30114	Copy down to Row 8
7	6.5	0.05399	0.12952	0.18265	0.15715
8	7.5	0.00443	0.01753	0.03174	0.08357
9	∏(Likelihood)	0.1425×10^{-6}	0.3108×10^{-6}	0.2502×10^{-6}	=product(E3:E8)

Probability densities only relate to probabilities in the context of a range of values, so it appears that the definition of likelihood we have been using is too strong. As we noted earlier, if the observed variate is a continuous variable, then the probability of each particular value would simply be infinitesimal. The common argument is that what we are really talking about is the product of the probability densities for our particular data values, but over a tiny range around this value (Edwards, 1972; Press et al., 1989). In practice, we just calculate the probability density for precise values and not a range of values. Edwards (1972) suggests we are effectively using a constant range of 1 about each data value, but this is hardly what would be called a tiny range.

There is no such conceptual problem with probability density functions for discrete statistical distributions (e.g., binomial) because the probability densities are exact probabilities. It is only with continuous pdfs that a problem of interpretation arises. While this may seem pedantic and only related to definitions, the source of likelihoods for continuous variates has concerned mathematical statisticians for decades, and a variety of solutions have been proposed (Edwards, 1972).

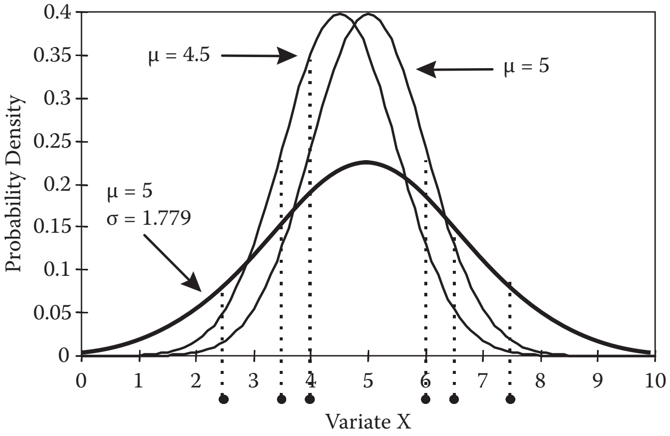

Fig ur e 3.11
Different hypothesized normal distributions superimposed upon the example's six data points. Shifting the mean away from 5.0 leads to a decrease in the overall likelihood. However, increasing the standard deviation from 1.0 to 1.779 increased the likelihood to its maximum. This implies the best fitting normal distribution is one with a mean of 5.0 and standard deviation of 1.779. Note that increasing the standard deviation widens the spread of the distribution, hence increasing the probability densities of the observations at 2.5, 7.5, 3.5, and 6.5, while decreasing the pdf values of the central observations, 4.0 and 6.0.

3.4.5 Maximum Likelihood Criterion

Earlier we estimated parameters using the minimal sum of the squared residual errors as a criterion of fit. Using maximum likelihood instead, parameter estimation is a matter of finding the set of parameters for which the observed data are most likely. In order to apply the maximum likelihood method to parameter estimation we need two things:

1. A list of hypotheses to be considered with the particular model (i.e., what combinations of parameters we are going to try)
2. A function required to calculate the probability density/likelihood of the observed data if the hypotheses were true

It is usual to search over ranges of parameter values, focusing with more detail on combinations with the largest likelihoods, just as with the nonlinear parameter searches already described and the likelihood search in Example Box 3.5.

3.4.6 Likelihoods with the Normal Probability Distribution

As already shown, probability densities for the normal distribution are calculated from the familiar

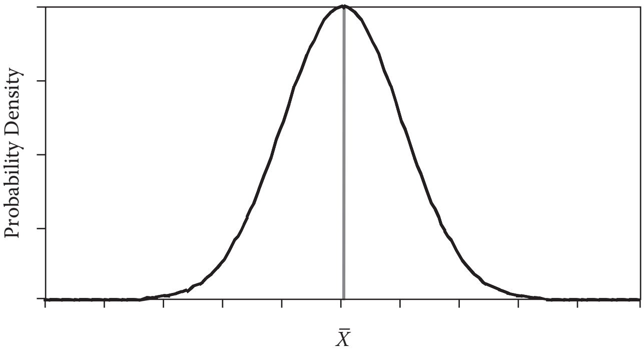

Fig ur e 3.12
Likelihood distribution for the standard normal function (mean = 0 and variance = 1). Note the symmetrical distribution of values around the mean. The tick marks on the X axis are in units of standard deviations above and below the mean. The height on the likelihood curve indicates the relative likelihood of randomly obtaining a particular X value. A value of X approximately two standard deviations from the mean is approximately one-twentieth as likely to occur as the mean value. Likelihoods have meaning only relative to one another.

$$L\left\{X|\mu,\sigma\right\} = \frac{1}{\sigma\sqrt{2\pi}}e^{\left(\frac{-(X-\mu)^2}{2\sigma^2}\right)} \tag{3.7}$$

where $L\{X|\mu,\sigma\}$ is the likelihood of any individual observation X, given, μ the population mean, and σ, the population standard deviation (that is, the likelihood of the data X, given the hypothesis about μ and σ; Figure 3.12). Using this equation with a wide range of X values produces the bell-shaped curve everyone associates with the normal distribution (Figure 3.12); the value of the likelihood is determined by the various possible combinations of mean and standard deviation values. The y axis represents the relative likelihood of observing each specific X value if one is sampling a given population. This probability distribution can be used to estimate likelihoods wherever the residual errors in a model are normally distributed.

As an example, we can consider the relationship between the caudal peduncle width and standard length of female orange roughy from the Lord Howe Rise. This linear relationship (Figure 3.13) can be described by a very simple model, a one-parameter linear regression (where C is the caudal peduncle width and S is the standard length):

$$C_i = bS_i + \varepsilon_i = bS_i + N\left(0,\sigma^2\right) \tag{3.8}$$

The residuals (ε_i) are assumed to be normally distributed with a mean of zero and variance σ^2, so if we use maximum likelihood methods, there are two parameters to estimate: b and σ. The probability density of any particular

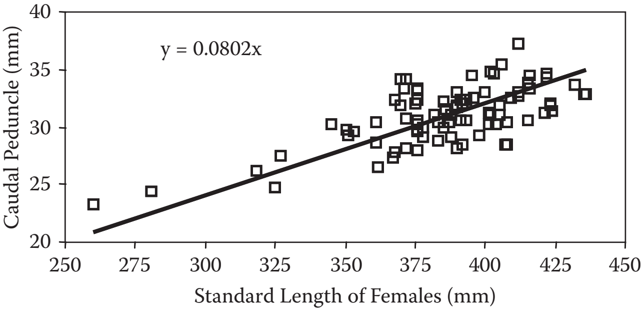

Fig ur e 3.13
Relation between caudal peduncle and standard length for female orange roughy from the Lord Howe Rise. The regression line has an $r^2 = 0.3855$, $F = 52.1$, $P < 0.0001$, df = 83. (Data described in Haddon, 1995.)

observation C_i, given values for b, σ, and the data S_i, is $L\{C_i|b, \sigma, S_i\}$ and can be obtained:

$$L\left\{C_i\middle|b,\sigma,S_i\right\} = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{-(C_i - bS_i)^2}{2\sigma^2}\right)} \tag{3.9}$$

The total probability density of all n observations given a particular pair of values for the parameters b and σ is just the product of the probability density for each of the n separate observations. Probability densities are calculated for independent events and, as such, have to be multiplied together and not added; hence, we do not use Σ (the sum) but instead ∏ (capital Pi) the product:

$$L\left\{C\middle|b,\sigma,S\right\} = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{-(C_i - bS_i)^2}{2\sigma^2}\right)} \tag{3.10}$$

As probability densities are commonly rather small numbers, the risk of rounding errors becomes great if many of them are multiplied together. So, given a function $f(X)$, we should remember that

$$\prod_{i=1}^{n} f\left(X_i\right) = e^{\sum Ln\left(f(X_i)\right)} \tag{3.11}$$

which simply says that the ∏, or product, of a series of values is the same as the antilog of the sum of the logs of those same values. If we omit the antilog we would be dealing with log-likelihoods. Using this latter approach, tiny numbers and the potential for rounding errors may be avoided. Equation 3.10 would take the form

$$LL\left\{C|b,\sigma,S\right\}=\sum_{i=1}^{n}Ln\left[\frac{1}{\hat{\sigma}\sqrt{2\pi}}e^{\left(\frac{-(C_i-bS_i)^2}{2\hat{\sigma}^2}\right)}\right] \tag{3.12}$$

Equation 3.12 can be simplified by expanding the logarithm and removing the terms that stay constant from the summation term as the parameters change.

$$LL=nLn\left(\frac{1}{\hat{\sigma}\sqrt{2\pi}}\right)+\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}\left[-\left[\left(C_i-bS_i\right)^2\right]\right] \tag{3.13}$$

where σ hat squared

$$\hat{\sigma}^2=\frac{\sum_{i=1}^{n}\left(C_i-bS_i\right)^2}{n} \tag{3.14}$$

is the variance estimated from the data (hence the hat). It is the maximum likelihood estimate so we divide by *n* and not *n* – 1. By expanding Equation 3.13, using Equation 3.14, we can produce a simplification that makes for easier calculation. The summation term in Equation 3.13 is cancelled by the inverse of $\hat{\sigma}^2$ from Equation 3.14, leaving –*n*/2. Further simplification is possible:

$$LL=nLn\left(\left(\hat{\sigma}\sqrt{2\pi}\right)^{-1}\right)-\frac{n}{2} \tag{3.15}$$

giving

$$LL=-n\left(Ln\left(2\pi^{\frac{1}{2}}\right)+Ln\left(\hat{\sigma}\right)\right)-\frac{n}{2} \tag{3.16}$$

and finally

$$LL=-\frac{n}{2}\left[Ln\left(2\pi\right)+2Ln\left(\hat{\sigma}\right)+1\right] \tag{3.17}$$

The objective when fitting data to a model is to maximize the log-likelihood, and any of Equations 3.12, 3.15, 3.16, or 3.17 could be used. Among people working with nonlinear models, it appears to be a tradition to minimize instead of maximizing a criterion of fit. In practice, all this means is that one minimizes the negative log-likelihood (i.e., for normal errors remove the leading negative symbol from Equations 3.16 and 3.17).

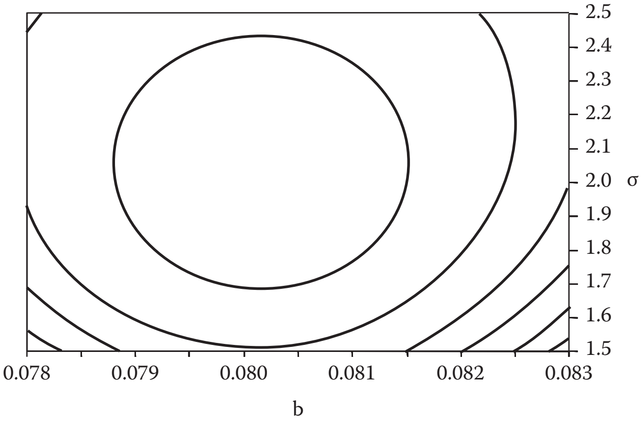

Fig ur e 3.14
Circular maximum likelihood contours for the female orange roughy from the Lord Howe Rise morphometric data (caudal peduncle vs. standard length) given different values of the single-parameter regression parameter *b* and the standard deviation of the residuals about the regression σ. Optimum fit at *b* = 0.08016 and σ = 1.990.

The log-likelihood, LL{C|*b*, σ, S}, can be back-transformed to the likelihood:

$$L\left\{C|, b, \sigma, S\right\} = e^{LL\{C|b,\sigma,S\}} \tag{3.18}$$

The optimum combination of *b* and σ can then be found by searching for the maximum of either the likelihood *L*{C|*b*,σ,*S*} or the log-likelihood (see Figure 3.14), or the minimum of the negative log-likelihood. In Excel, with this problem, the result obtained was the same irrespective of whether Equation 3.12 or Equation 3.17 was used. This occurred despite there being ninety-four pairs of data points. Generally, however, with large samples, Equation 3.17 would be the safest option for avoiding rounding errors and the machine limits to numbers (the likelihoods can become very small during the search).

There is no correlation between the parameter *b* and the standard deviation σ, as evidenced by the circular contours of the log-likelihood surface (Figure 3.14). The maximum likelihood produces the same estimate of *b* as the minimum sum of squared residual estimate (*b* = 0.080155).

3.4.7 equivalence with Least Squares

For linear and nonlinear models, having normally distributed residuals with constant variance, fitting the model by maximum likelihood is equivalent to ordinary least squares. This can be illustrated with a straight-line regression model:

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{3.19}$$

where the random deviations ε_i are independent values from a normal distribution with mean zero and variance σ^2 [$N(0, \sigma^2)$]. In other words, for a given value of the independent variable x_i, the response variable y_i follows a normal distribution with mean $\alpha + \beta x_i$ and variance σ^2, independently of the other responses. The probability distribution for the set of observations $y_1, y_2, \ldots, y_n$, is formed from the product of the corresponding normal pdf values (likelihoods), so that the likelihood of the parameters α and β is given by

$$L\left\{y \middle| \alpha,\beta,\sigma\right\} = \prod \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{-(y_i-\alpha-\beta x_i)^2}{2\sigma^2}\right)} \tag{3.20}$$

To fit by maximum likelihood we need to find the parameter values (α, β, and σ) that maximize the total likelihood (Equation 3.20). Equivalently, we could maximize the sum of the logarithm of the likelihoods (or minimize the sum of the negative logarithms).

$$LL\left\{y \middle| \alpha,\beta,\sigma\right\} = nLn\left(\sigma\sqrt{2\pi}\right)^{-1} + \frac{1}{2\sigma^2}\sum_{i=1}^{n} -\left(y_i - \alpha - \beta x_i\right)^2 \tag{3.21}$$

In Equation 3.21 the only part of the equation that is not constant as parameters alter and interact with the observed data is the summation term (the other terms being constant can validly be removed from the summation). The summation term is equivalent to the sum of squared residuals used in standard linear regression. Therefore, fitting by maximum likelihood will produce an equivalent result to fitting the line that minimizes the sum of the squared deviations of the observations about the fitted line, i.e., fitting by ordinary least squares. This is the case for all models as long as their residual errors are normal, additive, and with constant variance.

3.4.8 Fitting a Curve u sing Normal Likelihoods

The most commonly used equation to describe growth in fisheries modelling is still the von Bertalanffy (1938) growth curve:

$$\hat{L}_t = L_\infty\left(1 - e^{-K(t-t0)}\right) + \varepsilon \tag{3.22}$$

where L_t is the expected size at age t, L_∞ is the average maximum size, K is a growth rate parameter, $t0$ is the hypothetical age at zero length, and ε is the normal error term (see Chapter 9). Kimura (1980) provided a set of data relating to the growth of male Pacific hake. To fit the von Bertalanffy curve using normal likelihoods, we compare the observed length at age with those predicted from Equation 3.22 (Example Box 3.6, Figure 3.15).

EXAMPLE BOX 3.6

Fitting a von Bertalanffy growth curve using normal likelihoods. Data on male Pacific hake from Kimura (1980). In column C put Equation 3.22 =$B$1*(1–exp(–$B$2*(A6–$B$3))) and copy down to row 16. Sum column D in E1. We use the equivalent of Equation 3.17 in E4. The values in E3 and E4 will differ unless you put =E2 in B4. Why is that the case? Plot columns B as points and C as a line, against A (Figure 3.15). Note that altering the value for σ has no effect on the location or shape of the curve; it only affects the relative likelihoods. Find the optimum fit by using the solver to minimize the negative log-likelihood in either E3 or E4 by changing $\mathbf{L_\infty}$, K, and σ. Alternatively, solve by minimizing the sum of the squared residuals in E1. Do the results obtained with maximum likelihood differ from those obtained by least squares?

	A	B	C	D	E
1	$\mathbf{L_\infty}$	55.000		**SSQ**	=sum(D6:D16)
2	**K**	0.350		**St Dev ML**	=sqrt(E1/11)
3	**t0**	0.000		**negLL**	=–sum(E6:E16)
4	σ	1.000		**negLL** (eq 3.17)	=(11/2)*(Ln(2*PI())+2*Ln(E2)+1)
5	**Years**	**Length**	**Expect**	**Resid^2**	**Ln(Likelihood)**
6	1	15.4	16.242	=(B6–C6)^2	=Ln(normdist(B6,C6,$B$4,false))
7	2	26.93	27.688	=(B7–C7)^2	=Ln(normdist(B7,C7,$B$4,false))
8	3.3	42.23	37.672	=(B8–C8)^2	–11.3074
9	4.3	44.59	42.789	3.243425	–2.5407
10	5.3	47.63	46.395	1.525007	–1.6814
11	6.3	49.67	48.936	0.538431	–1.1882
12	7.3	50.87	50.727	0.020470	–0.9292
13	8.3	52.3	51.989	0.096835	–0.9674
14	9.3	54.77	52.878	3.579456	–2.7087
15	10.3	56.43	53.505	8.557435	–5.1977
16	11.3	55.88	53.946	3.739299	–2.7886

3.4.9 Likelihoods from the Lognormal Distribution

In fisheries models, the probability density function that is perhaps most commonly used to describe untransformed residual errors is the lognormal distribution. The distributions of catches, and of efforts, across a fleet are often lognormally distributed, while catch rates can usually be described using lognormal multiplicative errors (Figure 3.16). Events that relate to each other in an additive manner tend to be described by the normal distribution, while those that relate in a multiplicative way tend to be described by the lognormal distribution.

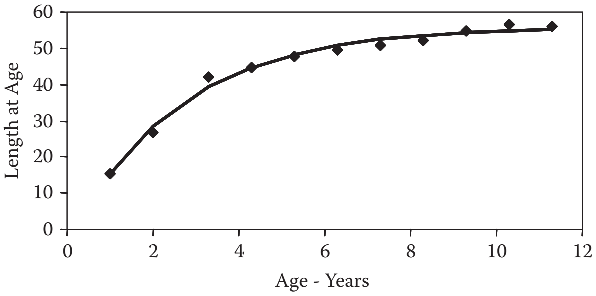

Figure 3.15
Kimura's (1980) growth data for male Pacific hake. The solid line represents the expected values (Example Box 3.6) using additive, normal random residual errors.

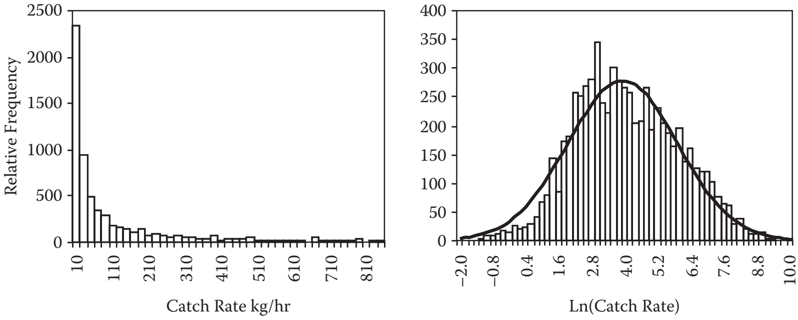

Figure 3.16
The left-hand panel shows the raw data from south coast catches of Blue Warehou (*Seriolella brama*) from the Australian Fisheries Management Authority's trawl database for the Australian South East Fishery. That this is lognormally distributed is illustrated in the right-hand panel, where the distribution is approximately normal with a mean (μ) of 3.93 and standard deviation (σ) of 2.0 after log transformation.

Natural logarithmic transformation of lognormally distributed data generates a normal distribution (multiplications are converted to additions). The probability density function for lognormal residual errors is (Hastings and Peacock, 1975)

$$L(x_i) = \frac{1}{x_i \sigma \sqrt{2\pi}} e^{\left[\frac{-(Ln(x_i)-Ln(m))^2}{2\sigma^2}\right]} \tag{3.23}$$

where $L(x_i)$ is the likelihood of the data point x_i in question, m is the median of the variable with $m = e^{\mu}$ (and $\mu = Ln(m)$), μ is estimated as the mean of $Ln(x_i)$, and σ is the standard deviation of $Ln(x_i)$. Equation 3.23 is equivalent to the pdf for normal distributions, with the data log-transformed and divided by each x_i value.

EXAMPLE BOX 3.7

Relationship between the normal pdf and the lognormal pdf. Column C is filled from C8:C1007, with numbers starting from 0.01 down to 10 in steps of 0.01. Column A is the natural log of column C, and column B is the normal likelihood (probability density) =normdist(A8,$D$4,$D$5,false). Column D is the lognormal likelihood (Equation 3.23), which is the normal likelihood in column B divided by the value of the observed data x in column C (i.e., =B8/C8 in D8). Column E uses the Excel function (e.g., =LogNormDist(C8, $D$4,$D$5) in E8), which provides the cumulative distribution by default. Notice that this function uses μ instead of the median m. Plot column B against A as one graph, and then column D against C, and column B against C as continuous lines. The first two should mimic Figure 3.17. Modify cells D3 and D5 and observe the effects on the shape and location of the different curves. Note the effect of division by x (column D). Note the value in D6. Why does it have this value (cf. Example Box 3.3)? Compare the value of D2 with the mode on the graph of column D; try putting 0.45 in D5. What is the impact on the sum in D6?

	A	B	C	D	E
1		Mean of the pdf	=D3*Exp((D5^2)/2)		
2		Mode of the pdf	=D3/(Exp(D5^2))		
3		Median m, of the pdf		**1**	
4		Ln(m) = μ =Avg(Ln(x))		=Ln(D3)	
5		σ = StDev(Ln(x))		**1**	
6		Sum of Likelihoods	=sum(D8:D1007)		
7	**Ln(x)**	**Normdist(μ,σ)**	**X**	**NormDist/x**	**Lognormdist**
8	=Ln(C8)	=normdist(A8,$d$4,$d$5,false)	0.01	=B8/C8	2.06279E–06
9	=Ln(C9)	=normdist(A9,$d$4,$d$5,false)	0.02	=B9/C9	4.57817E–05
10	–3.50656	0.000853	0.03	0.0284287	0.00022702
11	–3.21888	0.002244	0.04	0.0560987	0.00064353
12	–2.99573	0.004489	0.05	0.0897783	0.00136900
13	Copy these columns down to row 1007		0.06	Copy down to row 1007	

To use the normal pdf to generate probability density values for the lognormal distribution, one would log-transform the observed and expected values of the variable *x*, find the normal likelihood value, and then, strictly, divide this by the untransformed observed *x* value. This last step, of dividing by the observed *x* value, has no effect on the optimization of model fitting and is often omitted. Strictly, however, if one wanted to plot up the pdf of a lognormal distribution, Equation 3.23 should be used in full (Example Box 3.7, Figure 3.17). When describing the parameters of the lognormal distribution

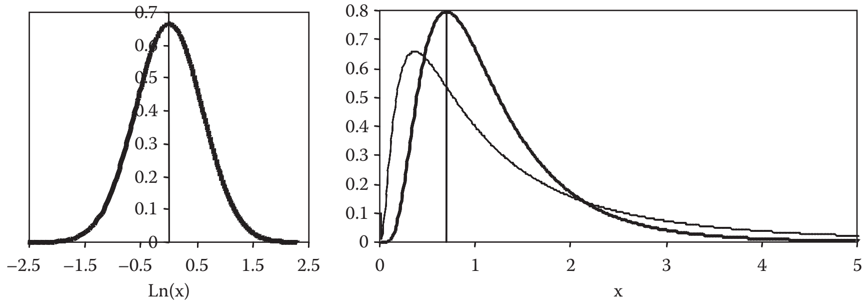

Fig ur e 3.17
Probability density distributions for the same data expressed as a lognormal distribution in the right-hand panel and as a normal distribution in the left-hand panel after the *x* variate was log-transformed. The thick line in both panels relates to data having a median (*m*) of 1 (i.e., $\mu = 0$) and a σ of 0.6. The thin vertical line in the right-hand graph indicates the mode of $0.6977 = 1/e^{0.36}$. The thin curve in the right-hand graph relates to parameters $m = 1$ and $\sigma = 1$. The *x* axes in both graphs have been truncated to areas where the probability density or likelihood values were significantly greater than zero (see Example Box 3.7). As the variance increases for a given median, the mode moves toward zero and the curve skews further to the right.

there is a potential for confusion, so care must be taken. The location parameter (position along the *x* axis) is the median ($m > 0$) of the *x* variate and not the mean, but the shape parameter σ ($\sigma > 0$) is the standard deviation of the log of the *x* variate. Given a continuous variate *x*, the main parameters of the lognormal distribution are (Hastings and Peacock, 1975):

1. The median:

$$m = e^{\mu} \tag{3.24}$$

   where μ is the mean of $Ln(x)$

2. The mode of the lognormal pdf:

$$m/e^{\sigma^2} = e^{(\mu - \sigma^2)} \tag{3.25}$$

   where σ is the standard deviation of $Ln(x)$

3. The mean of the lognormal pdf:

$$me^{(\sigma^2/2)} = e^{(\mu + \sigma^2/2)} \tag{3.26}$$

The lognormal distribution is always skewed to the right (Figure 3.17). The most obvious difference between the lognormal and normal curves is that the lognormal is always positive. In addition, as the mode is determined by both the median and the σ parameter (Equation 3.27), both can affect the location of the mode (Figure 3.17, Example Box 3.7).

The likelihood equation can be simplified as with the normal likelihoods. In fact, given the appropriate transformation, one can use the equivalent of Equation 3.17:

$$LL = -\frac{n}{2}\left[Ln(2\pi) + 2Ln(\hat{\sigma}) + 1\right] - \sum_{i=1}^{n} Ln(x_i) \tag{3.27}$$

where

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{\left(Ln(x_i) - Ln(\hat{x}_i)\right)^2}{n} \tag{3.28}$$

Note the maximum likelihood version of σ^2 using n instead of $n - 1$ (Example Box 3.5). The $\Sigma Ln(x)$ term is a constant and is usually omitted from calculations.

3.4.10 Fitting a Curve u sing Lognormal Likelihoods

Fitting a curve using lognormal residual errors is very similar to using normal random likelihoods. Recruitment in fisheries is notoriously variable, with occasional very large year classes occurring in some fisheries. Stock recruitment relationships are generally taken to exhibit lognormal residual errors. Numerous stock recruitment relationships have been described (see Chapter 10), but here we will restrict ourselves to the most commonly used equation, that by Beverton and Holt (1957). As an example, we will attempt to fit a stock recruitment curve to some real data. Hilborn and Walters (1992) indicate that there can be more than one form of the Beverton and Holt stock recruitment equation, but all would be expected to have lognormal residual errors. We will use the following version:

$$R_i = \frac{aS_i}{b + S_i} e^{N\left(0,\sigma^2\right)} \tag{3.29}$$

where R_{i} is the recruitment in year i, S_{i} is the spawning stock size that gave rise to R_{i}, a is the asymptotic maximum recruitment level, and b is the spawning stock size that gives rise to 50% of the maximum recruitment. The residual errors are lognormal with a μ of 0 and variance σ^2 (Example Box 3.8, Figure 3.18). Penn and Caputi (1986) provide data on the stock recruitment of Exmouth tiger prawns. They used a Ricker stock recruitment relationship, but we will use the Beverton and Holt relationship (Equation 3.29).

EXAMPLE BOX 3.8

Fitting a Beverton and Holt stock recruitment relationship using lognormal residual errors. The predicted recruitment values are in column C as =($B$1*A6)/($B$2+A6) and copied down. In column D put the log-likelihood =Ln(normdist(Ln(B6),Ln(C6),$B$3,false)). The natural logarithmic transformations are important; without them one would be using normal random errors and not lognormal. In E4, put Equation 3.17 =(C4/2)*(Ln(2*pi())+2*Ln(E1)+1). Plot column B against A (as points) and C against A as a line (cf. Figure 3.18). Use the solver and minimize E3 by varying cells B1:B3 to find the optimum line. If E4 is minimized instead, only cells B1:B2 need be varied (σ is estimated directly). If you want E3 and E4 to be the same, then put =E1 into B3. To generate the strict lognormal likelihoods, the normal likelihoods need to be divided by each observed recruitment value, i.e., =Ln(normdist(Ln(B6),Ln(C6),$B$3,false)/B6). While this alters the likelihoods generated, compare the results obtained when using this version to those obtained using the simpler version. Try minimizing the sum of squared residuals in cell E2. Are the results different from those obtained using maximum likelihood?

	A	B	C	D	E
1	**a**	27.366		**StDev ML**	=sqrt(E2/C4)
2	**b**	4.0049		**SSQ**	=sum(E6:E19)
3	σ	0.3519		**negLL**	=–sum(D6:D19)
4		**n**	=count(E6:E19)	**negLL** (eq 1.17)	5.2425
5	**Spawn**	**Recruit**	**Expect**	**LL**	**resid2**
6	2.4	11.6	10.25	0.0642	=(Ln(B6)–Ln(C6))^2
7	3.2	7.1	12.15	–1.0416	=(Ln(B7)–Ln(C7))^2
8	3.9	14.3	13.50	0.1122	=(Ln(B8)–Ln(C8))^2
9	5.7	19.1	16.07	0.0053	0.0298
10	6	12.4	16.41	–0.1917	0.0786
11	7.4	19.7	17.76	0.082	0.0108
12	8.2	37.5	18.39	–1.9258	0.5080
13	10	18.5	19.54	0.1134	0.0030
14	10.1	22.1	19.60	0.0671	0.0145
15	10.4	26.9	19.76	–0.259	0.0952
16	11.3	19.2	20.21	0.115	0.0026
17	12.8	21	20.84	0.1253	0.0001
18	18	9.9	22.39	–2.5626	0.6657
19	24	26.8	23.45	0.0537	0.0178

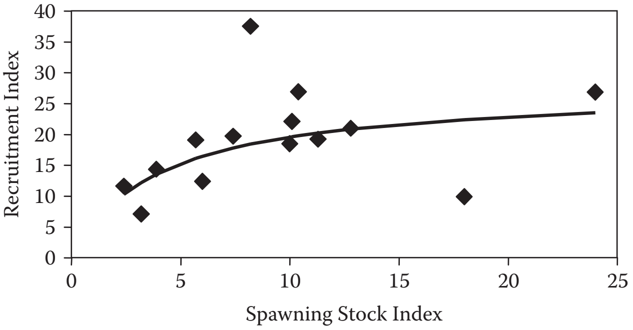

Fig ur e 3.18
A Beverton and Holt stock recruitment relationship (Equation 3.31) fitted to data from Penn and Caputi (1986) on Exmouth Gulf tiger prawns (*Penaeus semisulcatus*). The outliers to this relationship were thought to be brought about by extreme environmental conditions (see Chapter 10).

The estimated parameters obtained by using lognormal likelihoods are identical to those obtained through using either the minimum sum of squared residuals (on log-transformed data) or the normal likelihoods on log-transformed data (i.e., omitting the division by the observed recruitment, as with the x_i in Equation 3.23; see Example Box 3.8).

3.4.11 Likelihoods with the Binomial Distribution

Many texts introduce ideas relating to maximum likelihood estimation with a worked example using the binomial distribution. This seems to occur for two reasons. The first is that the binomial distribution can relate to very simple real examples, such as tagging-recapture experiments, where single parameters are to be estimated and only single likelihoods need be considered (i.e., no products of multiple likelihoods are required). The second is that the values calculated for this discrete distribution are true probabilities and not just probability densities. Thus, the complication of understanding likelihoods that are not true probabilities is avoided. Remember that this distinction is unnecessary with discrete probability distributions where the probability density function can only generate values (probabilities) for the possible discrete events. In fisheries stock assessment most analytical situations would need to use continuous probability density functions (pdfs), but there are situations where discrete pdfs, such as the binomial distribution, are necessary.

In situations where a study is determining whether an observation is true or false (a so-called Bernoulli trial; e.g., a captured fish either has or does not have a tag), and the probability of success is the parameter *p*, then it would generally be best to use the binomial distribution to describe observations. The binomial probability density function generates true probabilities and is characterized by two parameters, *n*, the number of trials, and *p*, the probability of success in a trial (an event proving to be true):

EXAMPLE BOX 3.9

Examining the properties of the binomial probability density function (Equation 3.30). The parameters of the binomial are n, the number of trials/observations, and p, the probability of a successful trial or observation. Set up a worksheet as below, where the column of m values (the number of observed successes) stretches from 0 to n down column A. Down column B insert and copy down =binomdist (A4, $B$1, $B$2, false) to obtain the likelihoods or probabilities directly (look up this function in the help). Down column C insert and copy down =fact($B$1)/(fact(A4)*fact($B$1–A4)) to obtain Equation 3.31. Put =($B$2^A4)*(1–$B$2)^($B$1–A4) into D4 and copy down, and then put =C4*D4 in E4 and copy down to obtain the likelihoods again. By plotting column B against column A as a simple scatterplot, you should be able to mimic Figure 3.19. Vary the value of p and observe how the distribution changes. Vary n (adjusting the length of the columns of numbers to match the n value) and see how above a value of 170 the binomdist() function operates beyond that of the fact() function. Look ahead to Equation 3.39 (the log-likelihood) and implement that on this worksheet to see how it handles the larger values of n.

	A	B	C	D	E
1	**n**	20			
2	**p**	0.25			
3	**m**	**Binomdist**	**FactTerm**	**p(1–p)term**	**Likelihood**
4	0	0.003171	1	0.003171	0.00317
5	1	0.021141	20	0.001057	0.02114
6	2	0.066948	190	0.000352	0.06695
7	3	0.133896	1140	0.000117	0.13390
8	4	0.189685	4845	3.92E-05	0.18969
	Extend down to equal n	Copy down to match column A			

$$P\left\{m \middle| n,p\right\} = \left[\frac{n!}{m!\left(n-m\right)!}\right] p^{m}\left(1-p\right)^{(n-m)} \tag{3.30}$$

which is read as the probability of *m* events proving to be true out of *n* trials (a sample of *n*), where *p* is the probability of an event being true (see Example Box 3.9). The term (1 – *p*) is often written as *q*, that is, (1 – *p*) = *q*. The "!" symbol denotes factorial. The term in the square brackets in Equation 3.30 is the number of combinations that can be formed from *n* items taken *m* at a time, and is sometimes written as

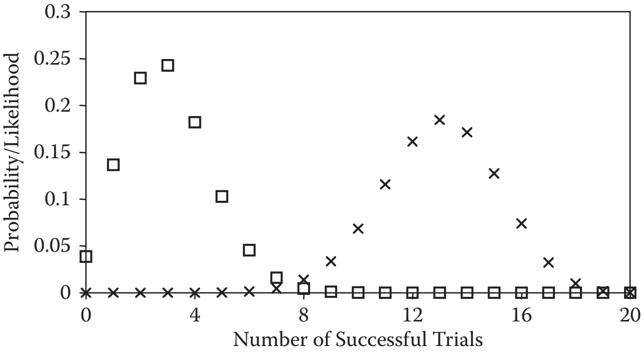

Fig ur e 3.19
Two examples of the binomial distribution. The left-hand set of squares are for n = 20 and p = 0.15, while the right-hand set of crosses are for n = 20 and p = 0.65. Note that zero may have a discrete probability, and that large p values tend to generate discrete, approximately normal distributions.

$$\left(\begin{array}{c} n \\ m \end{array} \right) = \frac{n!}{m!(n-m)!} \tag{3.31}$$

It is always the case that $n \geq m$ because one cannot have more successes than trials. The binomial distribution can vary in shape from highly skewed to approximately normal (Figure 3.19, Example Box 3.9).

Tagging programs, designed to estimate the size of a population, provide a common example of where a binomial distribution would be used in fisheries. Thus, if one has tagged a known number of animals, n_1, and later obtained a further sample of n animals from the population, it is possible to estimate the total population, assuming all of the assumptions of this form of tagging manipulation have been kept. The observation is of m, that is, how many in a sample of n are tagged (i.e., tagged = true, untagged = false).

A real example of such a study was made on the population of New Zealand fur seal pups (*Arctocephalus forsteri*) on the Open Bay Islands off the west coast of the South Island of New Zealand (York and Kozlof, 1987; Greaves, 1992). The New Zealand fur seal appears to be recovering after having been badly overexploited in the nineteenth century, with new haul-out sites starting to be found in the South and North Island. Exploitation officially ceased in 1894, with complete protection within the New Zealand Exclusive Economic Zone beginning in 1978 (Greaves, 1992). In cooperation with the New Zealand Department of Conservation, Greaves journeyed to and spent a week on one of these offshore islands. She marked 151 fur seal pups by clipping away a small patch of guard hairs on their heads, and then conducted a number of colony walk-throughs to resight tagged animals (Greaves, 1992). Each of these walk-throughs constituted a further sample of varying sizes, and differing numbers of animals

TABLe 3.1

Counts of New Zealand Fur Seal Pups Made by Greaves (1992) on Open Bay Island, West Coast, South Island, New Zealand

m	n	X	95%U	95%L	StErr
32	222	1,020	704	1,337	161.53
31	181	859	593	1,125	135.72
29	185	936	634	1,238	153.99

Note: She had tagged 151 animals (i.e., $n_1 = 151$). The column labelled n is the subsequent sample size, m is the number of tags resighted, X is the population size, and 95%L and 95%U are the lower and upper 95% confidence intervals, respectively, calculated as ±1.96 times StErr, the standard error. Calculations are as per Equations 3.33 and 3.34. The first two rows are individual independent samples, while the last row is the average of six separate counts (data from Greaves, 1992). The average counts lead to population estimates that are intermediate in value.

were found tagged in each sample (Table 3.1). The question is: What is the size of the fur seal pup population (X) on the island?

All the usual assumptions for tagging experiments are assumed to apply; i.e., we are dealing with a closed population—no immigration or emigration, with no natural or tagging mortality over the period of the experiment, no tags are lost, and tagging does not affect the recapture probability of the animals. Greaves (1992) estimated all of these effects and accounted for them in her analysis. Having tagged and resighted tags in a new sample, a deterministic answer can be found with the Peterson estimator (Caughley, 1977; Seber, 1982):

$$\frac{n_1}{X} = \frac{m}{n} \quad \therefore \quad X = \frac{n_1 n}{m} \tag{3.32}$$

where n_1 is the number of tags in the population, n is the subsequent sample size, m is the number of tags recaptured, and X is the population size. An alternative estimate adjusts the counts on the second sample to allow for the fact that in such cases we are dealing with discrete events. This is Bailey's adjustment (Caughley, 1977):

$$X = \frac{n_1(n+1)}{m+1} \tag{3.33}$$

Like all good estimators, it is possible to estimate the standard error of this estimate and thereby generate 95% confidence intervals around the estimated population size:

$$\text{StErr} = \sqrt{\frac{n_1^2(n+1)(n-m)}{(m+1)^2(m+2)}} \tag{3.34}$$

Instead of using the deterministic equations, a good alternative would be to use maximum likelihood to estimate the population size X, using the binomial probability density function. We will continue to refer to these as likelihoods even though they are also true probabilities.

We are only estimating a single parameter, X, the population size, and this entails searching for the population size that is most likely given the data. With the binomial distribution, $P\{m|n,p\}$, Equation 3.30 provides the probability of observing m tagged individuals from a sample of n from a population with proportion p tagged (Snedecor and Cochran, 1967; Hastings and Peacock, 1975). If we implemented this exact equation in a spreadsheet, we would quickly meet the limitations of computer arithmetic. For example, in Excel, one can use the =fact() function to calculate the factorial of numbers up to 170, but beyond that leads to a numerical overflow, as the result is too large to be represented as a normal real number. Fortunately, Excel provides a =binomdist() function that can operate with much larger numbers. It seems likely that the binomial probabilities have been implemented as log-likelihoods for the calculation and then back-transformed. Given

$$P\{m|n,p\} = \left[\frac{n!}{m!(n-m)!}\right] p^m (1-p)^{(n-m)} \tag{3.35}$$

log-transforming the component terms

$$Ln\left(p^m (1-p)^{(n-m)}\right) = mLn(p) + (n-m)Ln(1-p) \tag{3.36}$$

and

$$Ln\left[\frac{n!}{m!(n-m)!}\right] = Ln(n!) - \left(Ln(m!) + Ln((n-m)!)\right) \tag{3.37}$$

noting that

$$Ln(n!) = \sum_{i=1}^{n} Ln(i) \tag{3.38}$$

we obtain the log-likelihood

$$LL\{m\,|\,n,p\} = \sum_{i=1}^{n} Ln(i) - \left(\sum_{i=1}^{m} Ln(i) + \sum_{i=1}^{n-m} Ln(i)\right) + mLn(p) + (n-m)Ln(1-p) \tag{3.39}$$

The proportion p of fur seal pups that are marked is, in this case, $p = 151/X$, and with the first example in Table 3.1, n_1 is 151, and with the Bailey correction, n is 222 + 1, and m is 32 + 1 (see Equation 3.33). The maximum likelihood

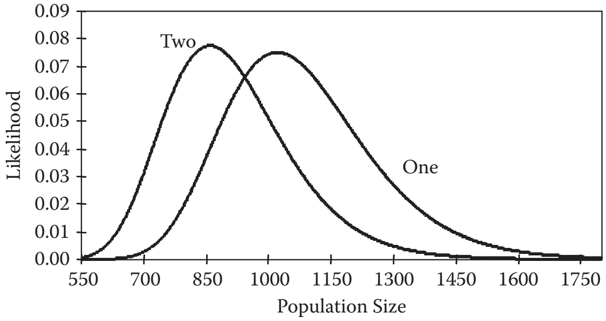

Fig ur e 3.20
Likelihood distribution against possible population size *X* for two estimates of fur seal pup population size made through a tagging experiment. The right-hand curve is that from 151 pups tagged, with a subsequent sample of 222 pups found to contain 32 tagged animals (with a mode of 1,020). The left-hand curve is for the same 151 tagged pups, with a subsequent sample of 181 pups containing 31 tagged animals (with a mode of 859). The modes are the same as determined by Equation 3.33 (Example Box 3.10).

estimate of the actual population size *X* is determined by searching for the value of *X* for which $P\{m|n,p\}$ is maximized. Using the likelihood equation (Equation 3.30), we can write

$$L\left\{data|X\right\}=\left[\frac{(223)!}{33!(223-33)!}\right]\left(\frac{151}{X}\right)^{33}\left(1-\frac{151}{X}\right)^{(223-33)} \tag{3.40}$$

Using Equation 3.35, and setting population size *X* to different values between 550 and 1,800, in steps of 1 (1,251 hypotheses), we gain a set of likelihoods to be plotted against their respective *X* values (outer curve in Figure 3.20). This can be repeated for the alternative set of observations from Table 3.1 (see Figure 3.20, Example Box 3.10).

3.4.12 Multiple Observations

When one has multiple surveys, observations, or samples, or different types of data, and these are independent of one another, it is possible to combine the estimates to improve the overall estimate. Just as with probabilities, the likelihood of a set of independent observations is the product of the likelihoods of the particular observations (Equation 3.41):

$$L\left\{O_1,O_2,....,O_n\right\}=L\left\{O_1\right\}\times L\left\{O_2\right\}\times......\times L\left\{O_n\right\} \tag{3.41}$$

This can be illustrated with the New Zealand fur seal pup tagging example (Example Box 3.10). The two independent resampling events, listed in Table 3.1, can be combined to improve the overall estimate. Instead of just taking the

EXAMPLE BOX 3.10

Use the binomial distribution to estimate population size and confidence intervals. Column A must be extended in steps of 1 down to row 1261 so that the 1,251 hypothesized population sizes can have their relative likelihood calculated. Put =binomdist(B$7+1,B$6+1,B$8/$A11,false) in column B and copy across into column C (note the $ symbols and their order). The +1s are the Bailey correction. The joint likelihoods are simply the separate values multiplied in column D, and these have been standardized to sum to 1 by dividing through by the sum of column D. Plot columns B and C against column A to obtain a graph akin to Figure 3.20. Vary the parameters in cells B6:C8 and consider how the likelihood profiles change. In the column next to the standardized likelihoods set out their cumulative distribution (i.e., in F11 put =E11, in F12 put =F11+E12, and copy down). Then search for the rows in which the values 0.025 and 0.975 occur. What do the population sizes at those cumulative likelihoods represent?

	A	B	C	D	E
1	**Experiment**	**1**	**2**		
2	**Deterministic**	1020.39	858.81		
3	**Determ StErr**	161.530	135.722		
4	**Upper 95%**	1336.99	=C2+1.96*C3		
5	**Lower 95%**	703.80	=C2–1.96*C3		
6	**Sample n**	222	181		
7	**Tags found m**	32	31		
8	**Tagged p**	151	151		
9	**Σ Likelihoods**	31.8084	=sum(C11:C1261)	1.1890	1
10	**Pop. Size**	**1**	**2**	**Joint**	**Std Joint**
11	550	2.906E–06	0.00058999	=B11*C11	=D11/$D$9
12	551	3.119E–06	0.00061706	=B12*C12	=D12/$D$9
13	552	3.345E–06	0.00064515	=B13*C13	=D13/$D$9
14	553	3.586E–06	0.00067430	Copy down	Copy down
15	Extend down	3.843E–06	0.00070453	2.707E–09	2.277E–09
16	To 1800 or to	4.116E–06	0.00073587	3.029E–09	2.548E–09
17	Row 1261	4.407E–06	0.00076835	3.386E–09	2.848E–09

average of the observations and putting those values through the deterministic equations (Table 3.1), the independent likelihood analyses can be combined using Equation 3.41. Thus, for each trial value of X, the separate likelihoods for each observation can be multiplied to give a joint likelihood (Figure 3.21, Example Box 3.10).

In this instance, where there were six separate sets of observations made, these could all, in theory, be combined to improve the overall estimate. In

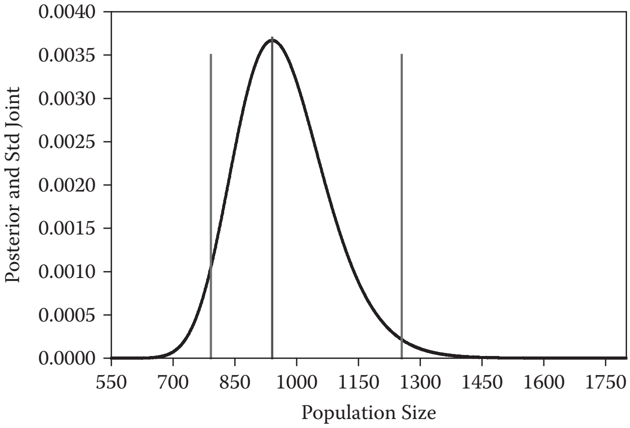

Fig ur e 3.21
The combined likelihoods vs. possible population size for the two sets of observations made on the tagged population of New Zealand fur seal pups. The two sets of likelihood values from Figure 3.20 were multiplied together at each hypothesized instance of population size to produce this composite likelihood distribution. Note the changed scale of the likelihood axis. Also shown are the 95% confidence intervals derived from the empirical likelihood profile. Note how the upper and lower intervals are not equal, as would be the case if we used the usual standard error approach (Example Box 3.10).

practice, it might be argued that all six samples were not strictly independent and so should not really be combined. As with all analyses, as long as the procedures are defensible, then the analyses can proceed (i.e., in this case it could be argued that the samples were independent—taken sufficiently far apart so there was no learning the locations of tagged pups, etc.).

3.4.13 Likelihoods from the Poisson Distribution

The Poisson distribution is another discrete statistical distribution whose probability density function generates actual probabilities. As with the binomial distribution, however, we will continue to refer to these as likelihoods.

The Poisson distribution is often used in ecology to represent random events. Most commonly, it will be used to describe the spatial distribution of organisms (if the mean density is roughly the same as the variance of the density, this is taken to suggest a random distribution) (Seber, 1982). It reflects one of the properties of the Poisson distribution, which is that the expectation of the distribution (its mean) is the same as its variance (Hastings and Peacock, 1975). A variate X describes the number of events per sample, and this can only assume values from 0 and upwards. To be distributed in a Poisson fashion, the variable should have two properties: it should be rare, that is, its mean value should be small relative to the number of events possible; and each event must be independent of the other events, that is, each

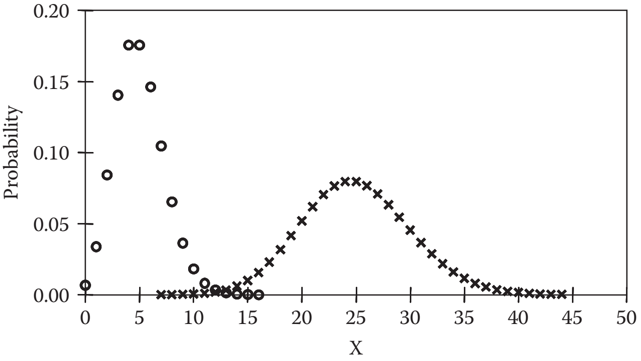

Fig ur e 3.22
Two examples of Poisson distributions. The left-hand distribution has a mean of 5, and the right-hand distribution a mean of 25. Note the top two values in each case (i.e., 4 and 5, and 24 and 25) have the same probabilities. The larger the value of μ, the closer the approximation is to a discrete normal distribution (Example Box 3.11).

event must be random with respect to each other. The Poisson probability density function has the following form:

$$P(X) = \frac{\mu^X}{e^{\mu} X!} \tag{3.42}$$

where X is the observed number of events and μ is the expected or mean number of events. As with the binomial distribution, the Poisson contains a factorial term, so the possibility of very small and very large numbers is a problem when computing the values of the distribution. The log transformation solves this problem (Figure 3.22, Example Box 3.11):

$$L\{X|\mu\} = \frac{\mu^X}{e^{\mu} X!} \tag{3.43}$$

Expanding the terms

$$LL\{X|\mu\} = Ln\left(\mu^X\right) - Ln\left(e^{\mu} X!\right) \tag{3.44}$$

and simplifying

$$LL\{X|\mu\} = X.Ln(\mu) - \mu - \sum_{i=1}^{X} Ln(i) \tag{3.45}$$

Hilborn and Walters (1992) present a hypothetical example about the analysis of a tagging-multiple-recapture study, where the Poisson distribution could be used in a fisheries context. We will use a similar example to illustrate the

EXAMPLE BOX 3.11

The properties of the Poisson distribution. We can compare the Excel Poisson function with the result of the log-likelihood calculation. Name cell B1 as mu. In column A put potential values of X from 0 to 50 down to row 53. In column B put Equation 3.45 and copy down; back-transform it in column C to obtain the likelihood or probability of the X value. In column D use the Excel Poisson function for comparison. Plot column C against A as points to obtain a graph like Figure 3.22. Vary the value of μ and observe the effect on the distribution of values. How does the spread or variance of the distribution change as μ increases?

	A	B	C	D
1	**μ**	4		
2	**X**	**Log-Likelihood**	**P(X)**	**P(X)**
3	0	=(A3*Ln(mu))–(mu+Ln(fact (A3)))	=exp(B3)	=poisson(A3,mu,false)
4	1	=(A4*Ln(mu))–(mu+Ln(fact (A4)))	=exp(B4)	=poisson(A4,mu,false)
5	2	–1.9206	0.1465	0.1465
6	3	–1.6329	0.1954	0.1954
7	4	–1.6329	0.1954	0.1954
8	5	–1.8560	0.1563	0.1563
9	6	–2.2615	0.1042	0.1042
10	Extend and copy these columns down to row 53 (where X = 50)			

importance of selecting the correct probability density function to represent the residual errors for a modelled situation.

The objective of the tagging-multiple-recapture analysis here is to estimate the constant rate of total mortality consistent with the rate of tag returns. Such surveys are a form of simplified Schnable census (Seber, 1982). It is simplified because while it is based around a multiple-recapture-tagging study, there is only a single tagging event. By tagging a known number of animals and tracking the numbers being returned in a set of equally spaced time periods (could be weeks, months, or years), an estimate of the total mortality experienced by the animals concerned can be determined. We will assume 200 animals were tagged and 57, 40, 28, 20, 14, 10, and 6 fish were recaptured. In total, twenty-five tags were therefore not retaken. The instantaneous total rate of mortality (the parameter to be estimated) is assumed to apply at a constant rate through the sampling period. As time passes, the number of fish alive in a population will be a function of the starting number and the number dying. As we saw in Chapter 2, this relationship can be represented as

$$N_t = N_0 e^{-Zt} \quad \text{or} \quad N_{t+1} = N_t e^{-Z} \tag{3.46}$$

where N_t is the number of fish alive at time t, N_0 is the number of fish alive at the start of observations, and Z is the instantaneous rate of total mortality. This is a straightforward exponential decline in numbers with time, where the numbers at time t are being multiplied by the survivorship to give the numbers at time $t + 1$. If tagging has no effect upon catchability, this relationship (Equation 3.46) can be used to determine the expected rate of return of the tags.

The expected number of tags captured at time t (C_t), given N_t and Z, is simply the difference between the expected number of tagged fish alive at time t and at time $t + 1$:

$$C_t = N_t - N_{t+1} = N_t - N_t e^{-Z} = N_t\left(1 - e^{-Z}\right) = \mu \tag{3.47}$$

Thus, μ is simply the number of tags multiplied by the complement of the survivorship between periods. The Poisson distribution can give the probability of capturing X tags, given μ. For example, with two hundred tags in the population and a $(1 - e^{-Z})$ of 0.05, we have $\mu = 200 \times 0.05 = 20$, and we would expect to recapture twenty tags in the first time period. However, we can calculate the likelihood of only capturing eighteen tags:

$$L\left\{18 \middle| \mu = 20\right\} = \frac{20^{18}}{e^{20}18!} = 0.08439 \tag{3.48}$$

By plotting likelihood/probability against hypothesized fishing mortality values in our example, a maximum can be seen to occur at approximately $Z = 0.35$ (Figure 3.23, Example Box 3.12). This plot also allows us to see that using a greater number of tags tends to increase the precision of the result (the spread of possible values becomes narrower).

It might seem to be a viable alternative to use a simple least squared residuals approach to match the number of tags observed against the predicted number of tags from Equation 3.47. Thus, we would be assuming normal random residual errors with a constant variance, and minimizing the sum of squared residuals should provide us with an optimum agreement between the observed number of tags and the expected:

$$SSQ = \sum\left(T_i - \left[N_t\left(1 - e^{-Z}\right)\right]\right)^2 \tag{3.49}$$

where T_i is the number of tags returned in period i. The difference between the two model fits relates to the different properties of the pdfs. With normal errors, the variance of the residuals is constant, but with Poisson errors, the variance increases with the expected number of tags. Strictly, with this model we should have used lognormal errors when using least squares; then the relationship with the classical catch curve of age against log numbers would become apparent and we would fit a regression to obtain

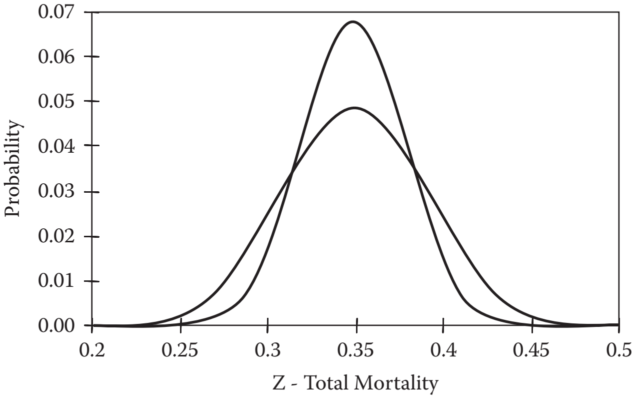

Fig ur e 3.23
Hypothesized total mortality vs. the likelihood for the imaginary example of the results from a tagging experiment. The total mortality consistent with the tagging results using Poisson likelihoods was approximately 0.348 (Example Box 3.12). The two curves refer to different tag numbers but with equivalent relative numbers of returns; more tags give better precision expressed as a narrower distribution of possibilities.

an estimate of *Z*. If it seems more likely that the variance will increase with number of tags returned, then clearly the Poisson is to be preferred. This must be decided independently of the quality of fit, because both analyses provide optimum but different parameter estimations (Example Box 3.12).

3.4.14 Likelihoods from the g amma Distribution

The gamma distribution is less well known to most ecologists than the statistical distributions we have considered in previous sections. Nevertheless, the gamma distribution is becoming more commonly used in fisheries modelling, especially in the context of length-based population modelling (Sullivan et al., 1990; Sullivan, 1992). The probability density function for the gamma distribution has two parameters, a scale parameter, b ($b > 0$; an alternative sometimes used is λ, where $\lambda = 1/b$), and a shape parameter, c ($c > 0$). The distribution extends over the range of $0 \le x \le \infty$, where x is the variate of interest. The expectation or mean of the distribution, $E(x)$, relates the two parameters, b and c. Thus,

$$E(x) = bc \quad \text{or} \quad c = \frac{E(x)}{b} \tag{3.50}$$

The variance of the distribution is b^2c, and for values of $c > 1$, the mode is calculated as $b(c-1)$, which is thus less than the expectation or mean (Figure 3.24).

A typical use of this distribution would be to describe the relative likelihood of each of a range of sizes to which a particular sized animal might

EXAMPLE BOX 3.12

A comparison of normal random residual errors and Poisson distributed residuals. Data from a hypothetical tagging program. The tags column records the number of tags returned over equal periods of time. Name cell C1 as Z, and cell F1 as N0. Fill columns D and E as shown. In C6 put the equivalent to Equation 3.45 =(B6*Ln(D6))–(D6+Ln(fact(B6))) and copy down. This contains the factorial function limited to a maximum X of 170. One could always use the Excel function =poisson(X, μ, false) to obtain the probabilities directly. Cell C3 is the negative sum of the log-likelihoods (=–sum(C6:C12)), and F3 the sum of the squared residuals (=sum(F6:F12)). Use the solver to minimize the negative log-likelihoods and the squared residuals in turn by changing Z (C1) and determine whether the results are the same. Plot column F against column D to observe the impact on the squared residuals of using the different residual structures. To observe the residuals proper, create a new column with =(B6-D6) in it and plot that against column D. With the least squares method the residuals are symmetrically and approximately equally arranged about the expected values. With the Poisson residuals there is an obvious trend with the expected value. Can you explain this in terms of the properties of the two statistical distributions? Put 400 in F1 and double each of the observed data values. Does the optimal solution change? What would be the advantage of tagging a larger sample at the start of the survey (cf. Figure 3.23)?

	A	B	C	D	E	F
1		**Z**	0.3479		**N0**	200
2						
3		**ΣLL**	17.103		**SSQ**	9.6696
4	**t**	**Tags**	**P(Tags)**	**$C_t = \mu$**	**Nt**	**$(\mu$-Tags$)^2$**
5	0				=N0*exp(–(Z*A5))	
6	1	57	–2.9686	=E5*(1–exp(–Z))	=N0*exp(–(Z*A6))	=(B6–D6)^2
7	2	40	–2.79279	=E6*(1–exp(–Z))	=N0*exp(–(Z*A7))	=(B7–D7)^2
8	3	28	–2.61751	29.30	70.4367	1.7026
9	4	20	–2.43277	20.69	49.7419	0.4828
10	5	14	–2.25752	14.61	35.1273	0.3777
11	6	10	–2.0836	10.32	24.8067	0.1028
12	7	6	–1.94994	7.29	17.5183	1.6599

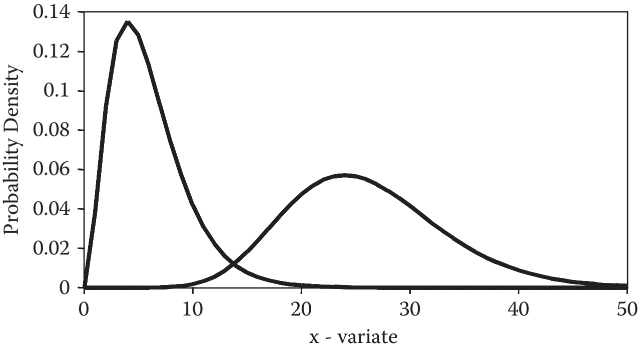

Fig ur e 3.24
Two examples of the gamma distribution with different parameter combinations. Both curves have a scale parameter, $b = 2$. The left-hand curve has an expected value of 6 (giving a shape parameter $c = 3$), while the right-hand curve has an expectation of 26 ($c = 13$). Note the modes are at $b(c - 1)$ and not at the expectations of the distributions (Example Box 3.14).

grow. The probability density function for determining the likelihoods for the gamma distribution is (Hastings and Peacock, 1975)

$$L\left\{x\middle|c,b\right\} = \frac{\left(\frac{x}{b}\right)^{(c-1)} e^{\frac{-x}{b}}}{b\Gamma(c)} \tag{3.51}$$

where x is the value of the variate, b is the scale parameter, c is the shape parameter, and $\Gamma(c)$ is the gamma function for the c parameter. Some books (and the Excel help file) give an alternative version:

$$L\left\{x\middle|c,b\right\} = \frac{x^{(c-1)} e^{-x/b}}{b^c\Gamma(c)} \tag{3.52}$$

but these are equivalent algebraically. Where the shape parameter, c, takes on integer values the distribution is also known as the Erlang distribution (Hastings and Peacock, 1975):

$$L\left\{x\middle|c,b\right\} = \frac{\left(\frac{x}{b}\right)^{(c-1)} e^{\frac{-x}{b}}}{b(c-1)!} \tag{3.53}$$

where the gamma function is replaced by factorial $(c - 1)$.

The gamma distribution is extremely flexible, ranging in shape from effectively an inverse curve, through a right-hand skewed curve, to approximately

EXAMPLE BOX 3.13

Likelihoods from the gamma and Erlang distributions. Continue the series of x values down to 50 (row 55), copying the respective column down. The Excel function in column B can give simple likelihoods or give the cumulative distribution function by using the "true" parameter instead of the "false" one. In this way, true probabilities can be derived. The Erlang distribution is identical to the gamma distribution when *c* is an integer; put =(((A5/$B$1)^($B$2–1))*exp(–A5/$B$1))/($B$1*fact($B$2–1)) into C5 and copy down. Plot column B against A to mimic Figure 3.24. Vary *b* and E(*x*) and determine the affect upon the curve. When *c* is not an integer, the fact function in the Erlang distribution truncates *c* to give an integer.

	A	**B**	**C**	**D**
1	**Shape b**	2	**Mode**	=B1*(B2–1)
2	**Scale c**	4	**Variance**	=(B1^2)*B2
3	**E(x)**	8		
4	**x**	**Gamma**	**Erlang**	
5	0.1	=gammadist(A5,$B$2,$B$1,false)	9.91E–06	
6	1	=gammadist(A6,$B$2,$B$1,false)	0.006318	
7	2	0.030657	0.030657	
8	3	0.062755	0.062755	
9	4	0.090224	0.090224	
10	Extend or copy down to row 55 where x = 50			

normal in shape (Figure 3.24, Example Box 3.13). Its flexibility makes it a very useful function for simulations (see Chapter 7).

It is possible that one might fit the gamma function to tagging data in order to provide a probabilistic description of how growth proceeds in a species (but one would need a great deal of tagging data; Punt et al., 1997a). In these instances of fitting the gamma distribution, the presence of the gamma function in the equation implies that it will be liable to numerical overflow errors within the numerical limits of the computer used. It would always be risk averse to work with log-likelihoods instead of likelihoods:

$$LL\left\{x|c,b\right\} = \left((c-1)Ln\left(\frac{x}{b}\right) - \frac{x}{b}\right) - \left(Ln(b) + Ln\left(\Gamma(c)\right)\right) \qquad (3.54)$$

This may appear to be rather of little assistance, as we are still left with the trouble of calculating the natural log of the gamma function. Surprisingly, however, this is relatively simple to do. Press et al. (1989) provide an excellent algorithm for those who wish to write their own procedure, and there is even a GammaLn function in Excel.

An example of using the gamma distribution in a real fisheries situation will be produced when we consider growth and its representation. It will be demonstrated when it is used to create the growth transition matrices used in length-based models.

3.4.15 Likelihoods from the Multinomial Distribution

We use the binomial distribution when we have situations where there can be two possible outcomes to an observation (true/false, tagged/untagged). However, there are many situations where there are going to be more than two possible discrete outcomes to any observation, and in these situations, we should use the multinomial distribution. In this multivariate sense, the multinomial distribution is an extension of the binomial distribution. The multinomial is another discrete distribution that provides distinct probabilities and not just likelihoods.

With the binomial distribution we used $P(m|n,p)$ to denote the likelihoods. With the multinomial, this needs to be extended so that instead of just two outcomes (one probability p), we have a probability for each of k possible outcomes (p_k) in our n trials. The probability density function for the multinomial distribution is (Hastings and Peacock, 1975)

$$P\left\{x_i|n,p_1,p_2,....,p_k\right\}=n!\prod_{i=1}^{k}\frac{\hat{p}_i^{x_i}}{x_i!} \tag{3.55}$$

where x_i is the number of times an event of type i occurs in n trials, n is the sample size or number of trials, and the p_i are the separate probabilities for each of the k types of events possible. The expectation of each type of event is $E(x_i) = np_i$, where n is the sample size and p_i the probability of event type i. Because of the presence of factorial terms that may lead to numerical overflow problems, a log-transformed version of Equation 3.55 tends to be used:

$$LL\left\{x_i|n,p_1,p_2,...,p_k\right\}=\sum_{j=1}^{n}Ln(j)+\sum_{i=1}^{k}\left[x_iLn(\hat{p}_i)-\sum_{j=1}^{x}Ln(j)\right] \tag{3.56}$$

In real situations the factorial terms will be constant and are usually omitted from the calculations; thus,

$$LL\left\{x_i|n,p_1,p_2,...,p_k\right\}=\sum_{i=1}^{k}\left[x_iLn(\hat{p}_i)\right] \tag{3.57}$$

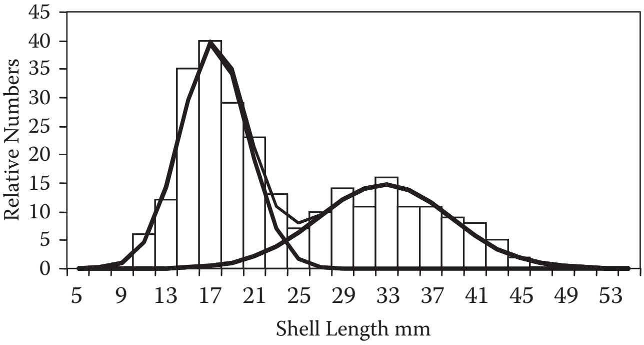

Figure 3.25
A subset of juvenile abalone data from Hope Island, Tasmania, taken in November 1992. Two modes are obvious and the solid line is the maximum likelihood best fit combination of two normal distributions (Example Box 3.14).

Examples may provide a better indication of the use of this distribution. Whenever we are considering situations where probabilities or proportions of different events or categories are being combined we should use the multinomial. This might happen in a catch-at-age stock assessment model that uses proportional catch-at-age. The model will predict the relative abundance of each age class, and these will be combined to produce sets of expected catch-at-age. The comparison of the observed with the expected proportions is often best done using the multinomial distribution. Another common use is in the decomposition of length frequency data into constituent modes (e.g., MacDonald and Pitcher, 1979; Fournier and Breen, 1983). We will consider the latter use and develop an example box to illustrate the ideas.

In November 1992, sampling began of juvenile abalone on Hope Island, near Hobart, Tasmania, with the aim of investigating juvenile abalone growth through the analysis of modal progression (Figure 3.25).

The main assumption behind the analysis of length frequency information is that observable modes in the data relate to distinct cohorts or settlements. Commonly a normal distribution is used to describe the expected relative frequency of each cohort (Figure 3.25), and these are combined to generate the expected relative frequency in each size class. The normal probability density function is commonly used to generate the expected relative proportions of each of the *k* observed size categories, for each of the *n* age classes. We subtract the cumulative probability for the bottom of each size class *i* from the cumulative probability of the top of each size class:

$$p_{S_k} = \int_{-\infty}^{TopS_i} \frac{1}{\sigma_n\sqrt{2\pi}} e^{\frac{-(S_i-\mu_n)^2}{2\sigma_n}} - \int_{-\infty}^{BotS_i} \frac{1}{\sigma_n\sqrt{2\pi}} e^{\frac{-(S_i-\mu_n)^2}{2\sigma_n}} \qquad (3.58)$$

where μ_n and σ_n are the mean and standard deviation of the normal distributions describing each cohort *n,* and S_i is the observed frequency of size or length class *i*. Alternative cumulative statistical distributions, such as the lognormal or gamma, could be used in place of the normal. As we are dealing with expected relative proportions and not expected relative numbers, it is not necessary to be concerned with expected numbers in each size class. However, it is suggested that it is best to have a graphical output, similar to Figure 3.25, in order to have a visual appreciation of the quality of fit. To obtain expected frequencies, it is necessary to constrain the total expected numbers to approximately the same as the numbers observed. The log-likelihoods from the multinomial are

$$LL\left\{S|\mu_n,\sigma_n\right\} = -\sum_{i=1}^{k} S_i Ln\left(\hat{p}_i\right) = -\sum_{i=1}^{k} S_i Ln\left(\frac{\hat{S}_i}{\sum \hat{S}_i}\right) \tag{3.59}$$

where the μ_n and σ_n are the mean and standard deviations of the *n* cohorts being considered. There are *k* length classes and S_i is the observed frequency of size or length class *i*, while p_i hat is the expected proportion of length class *i* from the combined normal distributions. Being the negative log-likelihood, the objective would be to minimize Equation 3.59 to find the optimum combination of the *n* normal distributions (cohorts; Example Box 3.14).

3.5 Bayes' Theorem

3.5.1 introduction

There has been a recent expansion in the use of Bayesian statistics in fisheries science (McAllister et al., 1994; McAllister and Ianelli, 1997; Punt and Hilborn, 1997; Chen and Fournier, 1999; see Dennis, 1996, for an opposing view). An excellent book relating to the use of these methods was produced by Gelman et al. (2004). Here we are not going to attempt a review of the methodology as it is used in fisheries; a detailed introduction can be found in Punt and Hilborn (1997), and there are many more recent examples. Instead, we will concentrate upon the foundation of Bayesian methods as used in fisheries and draw some comparisons with maximum likelihood methods.

Bayes' theorem is based around a manipulation of conditional probabilities. Thus, if an event, labelled *A* follows a number of possible events B_i, then we can develop Bayes' theorem by considering the probability of observing a particular B_i given that *A* has occurred:

$$P\left(B_i|A\right) = \frac{P(A \,\&\, B_i)}{P(A)} \tag{3.60}$$

EXAMPLE BOX 3.14

Using the multinomial to fit multiple normal distributions to length-frequency data, extend the series in column A in steps of two down to 55 in row 32, and similarly in column B. Extend the observed frequencies such that, as shown, in row 15 with the bottom of the size class at 21 and the top at 23, obs = 23, then Bot of 23 has obs = 13, then extending column F downwards, 7, 10, 14, 11, 16, 11, 11, 9, 8, 5, 2, and then zero for Bot of 47 to 55. Put =(normdist($B7,C$1,C$2,TRUE)-normdist($A7,C$1,C$2,TRUE))*C$3 into C7 (Equation 3.58 times N), and copy it into D7, and down to row 32. Row 5 contains the sum of each of the columns, as in C5, =sum(C7:C32). Plot column F as a histogram against column A. Add column E to this plot as a line chart (cf. Figure 3.25). The values in C1:D3 are initial guesses, the quality of which can be assessed visually by a consideration of the plotted observed vs. expected. Cell G2 contains (-sum(H7:H32))+(F5-E5)^2. The second term is a penalty term designed to force the sum of the expected frequencies to equal the observed total. This has no effect on the relative proportions of each cohort. Use the solver to minimize G2 by changing cells C1:D3. Try different starting points. Set up a new column that provides normal residuals, i.e., =(F7-E7)^2, sum, and minimize them (use the same penalty term to limit the relative frequencies). Are the optimum answers different? Compare the graphical images. Which do you prefer, that from multinomial or that from normal random residuals?

	A	B	C	D	E	F	G	H
1		**Mean**	18.3318	33.7929				
2		**Var**	2.9819	5.9919		**ML**	706.2087	
3		**N**	151.0604	110.9462				
4								
5		**Sum**	151.060	110.940	262	262		
6	**Bot**	**Top**	**Cohort1**	**Cohort2**	**Expt**	**Obs**	**P_L**	**LL**
7	5	7	0.01033	0.00035	=C7+D7	0	=E7/E$5	=F7*Ln(G7)
8	7	9	0.12135	0.00152	=C8+D8	0	=E8/E$5	=F8*Ln(G8)
9	9	11	0.92078	0.00595	0.92673	0	0.00354	0.00000
10	11	13	4.51878	0.02096	4.53973	6	0.01733	–24.33285
11	13	15	14.35710	0.06603	14.42313	12	0.05505	–34.79413
12	15	17	29.55476	0.18633	29.74109	35	0.11352	–76.15351
13	17	19	39.43916	0.47081	39.90997	40	0.15233	–75.26872
14	19	21	34.12401	1.06524	35.18926	29	0.13431	–58.22050
15	21	23	19.14159	2.15824	21.29983	23	0.08130	–57.72184
16	23	25	6.95823	3.91567	10.87390	13	0.04150	–41.36572

In an analogous fashion, we can consider the conditional probability of the event A given a particular B_i:

$$P\left(A|B_i\right)=\frac{P\left(A \,\&\, B_i\right)}{P\left(B_i\right)} \tag{3.61}$$

Rearranging Equation 3.61,

$$P\left(A|B_i\right)P\left(B_i\right)=P\left(A \,\&\, B_i\right) \tag{3.62}$$

Substituting Equation 3.62 into Equation 3.60, we obtain the basis of Bayes' theorem:

$$P\left(B_i|A\right)=\frac{P\left(A|B_i\right)P\left(B_i\right)}{P\left(A\right)} \tag{3.63}$$

If we translate the A as the data observed from nature and the B_i as the separate hypotheses (as models plus parameter values), we can derive the form of Bayes' theorem as it is used in fisheries. The $P(A|B_i)$ is just the likelihood of the data A given the hypothesis B_i. The $P(B_i)$ is the probability of the hypothesis before any analysis or consideration of the data. This is known as the prior probability of the hypothesis B_i. The $P(A)$ is simply the combined probability of all the combinations of data and hypotheses, which is why this works so well for closed systems such as card games and other constrained games of chance.

$$P\left(A\right)=\sum_{i=1}^{n}P\left(A|B_i\right)P\left(B_i\right) \tag{3.64}$$

3.5.2 Bayes' Theorem

As stated earlier, Bayes' theorem relates to conditional probabilities (Gelman et al., 2004), so that when we are attempting to determine which of a series of n discrete hypotheses is most probable, we use

$$P\left\{H_i|data\right\}=\frac{L\left\{data|H_i\right\}P\left\{H_i\right\}}{\sum_{i=1}^{n}\left[L\left\{data|H_i\right\}P\left\{H_i\right\}\right]} \tag{3.65}$$

where H_i refers to hypothesis i out of the n being considered (a hypothesis would be a particular model with a particular set of parameter values) and the data are just the data to which the model is being fitted. $P\{H_i|data\}$ is the *posterior* probability of the hypothesis (a strict probability between 0 and 1) given the data (and any prior information). $P\{H_i\}$ is the *prior* probability of the hypothesis before the observed data are considered; once again, this is a strict probability where the sum of the priors for all hypotheses being considered must be 1. Finally, $L\{data|H_i\}$ is the likelihood of the data given hypothesis *i*, just as previously discussed in the maximum likelihood section (analogous, for example, to Equations 3.7 and 3.35). If the parameters are continuous variates (e.g., L_∞ and K from the von Bertalanffy curve), alternative hypotheses have to be described using a vector of continuous parameters instead of a list of discrete parameter sets, and the Bayesian conditional probability becomes continuous:

$$P\left\{H_i|data\right\} = \frac{L\left\{data|H_i\right\}P\left\{H_i\right\}}{\int L\left\{data|H_i\right\}P\left\{H_i\right\}dH_i} \tag{3.66}$$

In fisheries, to use Bayes' theorem to generate the required posterior distribution we need three things:

1. A list of hypotheses to be considered with the model under consideration (i.e., the combinations of parameters and models we are going to try)
2. A likelihood function required to calculate the probability density of the observed data given each hypothesis *i*, $L\{data|H_i\}$
3. A prior probability for each hypothesis, normalized so that the sum of all prior probabilities is equal to 1.0

Apart from the requirement for a set of prior probabilities, this is identical to the requirements for determining the maximum likelihood. The introduction of prior probabilities is, however, a big difference, and is something we will focus on in our discussion.

If there are many parameters being estimated in the model, the integration involved in determining the posterior probability in a particular problem can involve an enormous amount of computer time. There are a number of techniques used to determine the Bayesian posterior distribution, and Gelman et al. (2004) introduce the more commonly used approaches. We will introduce and discuss one flexible approach to estimating the Bayesian posterior probability in Chapter 8, dealing with the characterization of uncertainty. This is effectively a new method for model fitting, but for convenience will be included in the section on uncertainty. The explicit objective of a Bayesian analysis is not just to discover the mode of the posterior, which in maximum likelihood terms might be thought to represent the optimum model. Rather, the aim is to explicitly characterize the relative probability of the different possible outcomes from an analysis, that is, to characterize the uncertainty about each parameter and model output. There may be a most probable result, but it is presented in the context of the distribution of probabilities for all other possibilities.

3.5.3 Prior Probabilities

There are no constraints placed on how prior probabilities are determined. One may already have good estimates of a model's parameters from previous work on the same or a different stock of the same species, or at least have useful constraints on parameters (such as negative growth not being possible or survivorship > 1 being impossible). If there is insufficient information to produce informative prior probabilities, then commonly a set of uniform or noninformative priors are adopted in which all hypotheses being considered are assigned equal prior probabilities. This has the effect of assigning each hypothesis an equal weight before analysis. Of course, if a particular hypothesis is not considered in the analysis, this is the same as assigning that hypothesis (model plus particular parameters) a weighting or prior probability of zero.

One reason why the idea of using prior probabilities is so attractive is that it is counterintuitive to think of all possible parameter values being equally likely. Any experience in fisheries and biology provides one with prior knowledge about the natural constraints on living organisms. Thus, for example, even before thorough sampling it should have been expected that a deep-water (>800 m depth) fish species, like orange roughy (*Hoplostethus atlanticus*), would likely be long-lived and slow growing. This characterization is a reflection of the implications of living in a low-temperature and low-productivity environment. One of the great advantages of the Bayesian approach is that it permits one to move away from the counterintuitive assumption of all possibilities being equally likely. One can attempt to capture the relative likelihood of different values for the various parameters in a model in a prior distribution. In this way, prior knowledge can be directly included in analyses.

Where this use of prior information can lead to controversy is when moves are made to include opinions. For example, the potential exists for canvassing a gathering of stakeholders in a fishery for their belief on the state of such parameters as current biomass (perhaps relative to five years previously). Such a committee-based prior probability distribution for a parameter could be included into a Bayesian analysis as easily as could the results of a previous assessment. There is often debate about whether priors from such disparate sources should be equally acceptable in a formal analysis. In a discussion on the problem of justifying the origin of priors, Punt and Hilborn (1997, p. 43) state:

> We therefore strongly recommend that whenever a Bayesian assessment is conducted, considerable care should be taken to document fully the basis for the various prior distributions…. Care should be taken when selecting the functional form for a prior because poor choices can lead to incorrect inferences. We have also noticed a tendency to underestimate uncertainty, and hence to specify unrealistically informative priors—this tendency should be explicitly acknowledged and avoided.

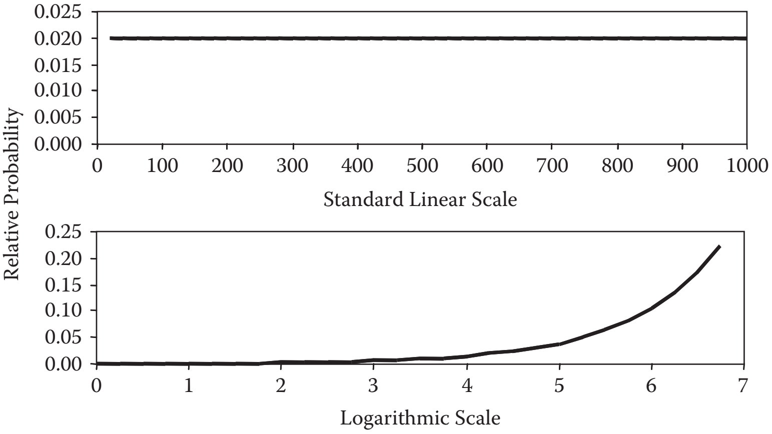

Fig ur e 3.26
The same data plotted on a linear scale (upper panel) and a natural logarithmic scale (lower panel). The uniform distribution on the linear scale is distorted when perceived in logarithmic space. Note the vertical scale in the log panel is an order of magnitude greater than on the linear scale.

The debate over the validity of using informative priors has been such that Walters and Ludwig (1994) recommended that noninformative priors be used as a default in Bayesian stock assessments. However, besides disagreeing with Walters and Ludwig, Punt and Hilborn (1997) highlighted a problem with our ability to generate noninformative priors (Box and Tiao, 1973). The problem with generating noninformative priors is that they are sensitive to the particular measurement system. Thus, a prior that is uniform on a linear scale will not appear linear on a log scale (Figure 3.26).

As fisheries models tend to be full of nonlinear relationships, the use of noninformative priors is controversial because a prior that is noninformative with respect to some parameters will most likely be informative toward others. While such influences may be unintentional, they cannot be ignored. The implication is that information may be included into a model in a completely unintentional manner, which is one source of controversy when discussing prior probabilities. If priors are to be used, then Punt and Hilborn's (1997) exhortation to fully document their origin and properties is extremely sensible advice.

3.5.4 An example of a u seful informative Prior

We already have experience with a form of Bayesian analysis when we were considering binomial likelihoods (see Example Box 3.10). We took two likelihood profiles of a fur seal pup population estimate, each made from 1,251 separate likelihoods, multiplied the respective likelihoods for each of 1,251 hypothesized population sizes, and standardized the results to sum to 1 in order to obtain approximate percentile confidence intervals. This algorithm may be represented as

$$P\left\{H_i\middle|data\right\}=\frac{L_1\left\{data\middle|H_i\right\}L_2\left\{data\middle|H_i\right\}}{\sum_{i=1}^{1251}\left[L_1\left\{data\middle|H_i\right\}L_2\left\{data\middle|H_i\right\}\right]} \tag{3.67}$$

where L_i refers to the likelihoods for the *i*th population estimation and the H_i refer to the 1,251 separate hypothesized population sizes. If we had standardized the first population estimate's likelihood profile to sum to 1, we could then have treated that as a prior on the population size. This could then have been used in Equation 3.65, along with the separate likelihoods for the second population estimate, and a formal posterior distribution for the overall population estimate been derived. Although these two algorithms sound rather different, they are, in fact, equivalent, as can be seen algebraically. Treating the first estimate as a prior entails standardizing each hypothesized population size's likelihood to sum to 1:

$$P\left\{H_i\right\}=\frac{L\left\{data\middle|H_i\right\}}{\sum_{i=1}^{n}L\left\{data\middle|H_i\right\}} \tag{3.68}$$

where n is the number of separate hypotheses being compared (1,251 in Example Box 3.10; see Example Box 3.15). The denominator in Equation 3.68 is, of course, a constant. Thus, expanding Equation 3.65 using Equation 3.68 we obtain

$$P\left\{H_i\middle|data\right\}=\frac{L_2\left\{data\middle|H_i\right\}\dfrac{L_1\left\{data\middle|H_i\right\}}{\sum_{i=1}^{n}L_1\left\{data\middle|H_i\right\}}}{\sum_{i=1}^{n}\left[L_2\left\{data\middle|H_i\right\}\dfrac{L_1\left\{data\middle|H_i\right\}}{\sum_{i=1}^{n}L_1\left\{data\middle|H_i\right\}}\right]} \tag{3.69}$$

As the sum of separate likelihoods term is a constant, it can be moved out of the denominator's overall summation term and we can shift the numerator's inverse sum of likelihoods below the divisor. When the two are brought together, they can thus cancel out:

$$=\frac{L_2\left\{data\middle|H_i\right\}L_1\left\{data\middle|H_i\right\}}{\dfrac{\sum_{i=1}^{n}L_1\left\{data\middle|H_i\right\}}{\sum_{i=1}^{n}L_1\left\{data\middle|H_i\right\}}\sum_{i=1}^{n}\left[L_2\left\{data\middle|H_i\right\}L_1\left\{data\middle|H_i\right\}\right]} \tag{3.70}$$

EXAMPLE BOX 3.15

Bayesian posteriors are equal to standardized likelihoods when the prior probability is uniform. Recover Example Box 3.10 and extend it to include a column (perhaps in G or H) in which the likelihoods for experiment 1 are standardized. Thus, in G11 put =B11/$B$9, and copy down. Use this as a prior by generating another column of numbers, so in H11 put =G11*C11, and copy down. Sum that column into H9. Finally, in column I (cell I11) put =H11/$H$9, and copy down. Column I would then contain the posterior distribution according to Equation 3.65. Plot this against column A and compare the curve with that found in column E (the standardized joint distribution). How similar are to two curves?

which is identical to Equation 3.67. Thus, the combination of two likelihood profiles from two separate assessments can be considered as an acceptable form of Bayesian analysis; that is, the origin of the prior distribution is not problematical.

With discrete statistical distributions, there is no problem with combining the separate likelihoods because they are probabilities and not just likelihoods. With the continuous variables found in most fisheries models, numerical methods would be needed to conduct the integration required, and this means that the answers would only be approximate. Nevertheless, the approximation would be acceptable assuming the integral step was sufficiently fine.

3.5.5 Noninformative Priors

If there is no earlier assessment and noninformative prior probabilities are to be used, then an even simpler adjustment can be made to Bayes' theorem. In the case of truly noninformative priors, the prior probability for each hypothesis would be equal or constant for each hypothesis, and thus the prior probability term could be removed from the summation term in the denominator:

$$P\left\{H_i \middle| data\right\} = \frac{P\left\{H_i\right\} P\left\{data \middle| H_i\right\}}{P\left\{H_i\right\} \sum_{i=1}^{n} P\left\{data \middle| H_i\right\}} \tag{3.71}$$

whereupon the prior terms can be cancelled and we are left simply with the standardized likelihoods (each separate likelihood divided by the sum of the likelihoods). Thus, using truly noninformative priors is the same as not using priors at all. The simplest approach then would be to omit them from the analysis altogether. Walters and Ludwig (1994) also point this out, but then proceed to discuss the problem of the difficulty of generating noninformative priors, and so do not take this further.

This all raises the importance of the problem of the apparent impossibility of generating priors that are noninformative over all parameters and model outputs. For a particular model, if generating a truly noninformative prior across all parameters and model outputs requires a great deal of work, or it even appears to be impossible, then trying to use noninformative priors would actually risk generating a biased analysis. If these interactions between prior distributions for different parameters and the effects of nonlinearity within the model act to distort the priors so that they are no longer truly noninformative, then we are left in a quandary. What this would imply is that even if we restrict ourselves to a straightforward maximum likelihood analysis as in Equation 3.70 (equivalent to using uniform priors in the linear scale), then we are really imputing informative priors in some parts of the model. This unwitting imputation of potential bias is startling and raises many questions about the optimum approach to any nonlinear analyses. It appears to imply that it is impossible to conduct a standardization of a full set of likelihoods without implying an unknown set of informative prior probabilities.

In opposition to this idea of the automatic imputation of informative priors, it appears that if priors are simply omitted from the analysis, then, from Equation 3.71, the omitted priors must have been, by definition, noninformative (on all scales) and somehow equal on all scales. This is a matter for more formal investigation, but its importance is crucial for the wide acceptance and further developments in the use of Bayesian analysis in fisheries modelling.

3.6 Concluding Remarks

Fisheries modellers can be an argumentative crowd, and each seems to have developed a set of methods that they favour. As seen above, there does not appear to be a criterion of quality of model fit that has no associated problems. Claims made that identify a particular strategy of analysis as being the optimum or best practice should therefore be contemplated with some doubt. One can use either maximum likelihood methods or Bayesian methods to generate assessments giving similar forms of output. At least some of the expressed enthusiasm for Bayesian methods appears to be excessive.

The optimum method to use in any situation depends largely on the objectives of the analysis. If all one wants to do is to find the optimum fit to a model, then it does not really matter whether one uses least squares, maximum likelihood, or Bayesian methods. Sometimes it can be easier to fit a model using least squares and then progress to using likelihoods or Bayesian methods to create confidence intervals and risk assessments.

Confidence intervals around model parameters and outputs can be generated using traditional asymptotic methods (guaranteed symmetric and, with strongly nonlinear models, only roughly approximate), using likelihood profiles or by integrating Bayesian posteriors (the two are obviously strongly related), or one can use bootstrapping or Monte Carlo techniques.

It is not the case that the more detailed areas of risk assessment are only possible using Bayesian methods. Bootstrapping and Monte Carlo methods provide the necessary tools with which to conduct such work. The primary concern should be to define the objective of the analysis. It would be bad practice to fit a model and not give some idea of the uncertainty surrounding each parameter and the sensitivity of the model's dynamics of the various input parameters.

Because there is no clear winner among the methodological approaches, if one has the time, it is a reasonable idea to use more than one approach (especially a comparison of likelihood profiles, Bayesian posteriors, and bootstrapping). If significant differences are found, then it would be well to investigate the reasons behind them. If different procedures suggest significantly different answers, it could be that too much is being asked of the available data and different analyses would be warranted.