4.2.2.1. *Method of Moments*

If the parameter θ can be expressed in terms of population means, variances, and covariances, and no other information is available beyond the sample values, a common procedure (though certainly not the only one) is "method-of-moments" estimation. In this case the population moments are estimated with the sample moments, e.g.,

$$\hat{\mu}_i(\theta) = \bar{x}_i$$

$$\hat{\sigma}_i(\theta)^2 = s_i^2$$

$$\hat{\sigma}_{ij}(\theta) = s_{ij},$$

with the subscript indicating that multiple attributes are recorded for a sample. The value of θ for which one or more of these equalities hold is the method-of-moments estimate of θ. Thus the functions $\mu_i(\theta)$, $\sigma^2{}_i(\theta)$, and/or $\sigma_{ij}(\theta)$ essentially determine the estimate $\hat{\theta}$. The "knowns" in this approach are sample values $x_{i1}, \ldots, x_{in}$, and the unknown is θ. A value for θ is to be chosen so that the population moments are equal to the sample moments.

Example

You wish to estimate the annual survival rate S and recovery rate f for mallards banded on their breeding grounds in the preseason and subjected to harvest during the ensuing two hunting seasons. By recovery rate is meant the probability that a banded bird is harvested and the band is reported to a central repository. A proportion $p_1 = f$ of bands is expected to be recovered in the year of banding. Because birds must survive until year 2 in order to be harvested that year, the proportion of bands expected to be recovered in the second year is $p_2 = Sf$ (assuming the same recovery rate in both years). If 1000 birds are banded in the preseason of year 1 and 90 and 60 bands are recovered in years 1 and 2, respectively, an estimate of S and f can be obtained with the method of moments by

$$\begin{aligned}\hat{p}_1(S, f) &= \hat{f}\\ &= 90/1000\\ &= 0.09\end{aligned}$$

and

$$\begin{aligned}\hat{p}_2(S, f) &= \widehat{Sf}\\ &= 60/1000\\ &= 0.06,\end{aligned}$$

from which we get $\hat{f} = 0.09$ and $\hat{S} = 0.67$. Note that the estimates of f and S are derived without any information about the underlying distribution of the recovery data, beyond the mathematical assumptions about the return proportions. Thus, the information necessary to assess the reliability of the estimates and to perform tests of statistical significance is missing.

4.2.2.2. *Maximum Likelihood Estimation*

A procedure used throughout this book to estimate parameters is maximum likelihood estimation. The procedure requires knowledge of the underlying distribution of a random sample, as well as the actual sample values. Thus, one presumes to know the mathematical form of the distribution function $f(\underline{x}|\theta)$, but not to know the value θ for the distribution. Under these conditions one estimates θ by sampling the population and using the distribution function as a likelihood function.

The difference between maximum likelihood estimation and other estimation techniques concerns the use of the likelihood function. Instead of choosing $\hat{\theta}$ to equate sample and population moments, for example, here we choose $\hat{\theta}$ to maximize $f(\underline{x}|\theta)$. Note, however, that the roles of parameter and variable have been switched: we wish to choose a value for the "variable" θ to maximize $f(\underline{x}|\theta)$, which is "parameterized" by the sample value $\underline{x}$. It is in this sense that the distribution function becomes a likelihood function. The usual convention is to express the likelihood function as $L(\theta|\underline{x})$ to emphasize the changed roles of θ and $\underline{x}$.

The usual method for determining maximum likelihood estimates is based on differentiation of the likelihood function. Assume for now that the $\hat{\theta}$ is the maximum likelihood estimator of a single parameter θ. Under certain mild conditions it can be shown that the value $\hat{\theta}$ maximizing $L(\theta|\underline{x})$ can be obtained by differentiating $\log(L)$ with respect to θ and setting the derivative to 0:

$$\frac{d[\log L(\theta|\underline{x})]}{d\theta} = 0.$$

The value of θ that satisfies this equation (which is known as the likelihood equation) also maximizes L and therefore is the maximum likelihood estimator (conditional on $\underline{x}$). In the case in which there are, say, k parameters in $\underline{\theta}$, then k likelihood equations are defined by partial differentiation of the log likelihood with respect to each parameter:

$$\frac{\partial[\log L(\underline{\theta}|\underline{x})]}{\partial\theta_i} = 0,$$

$i = 1, ..., k$. Simultaneous solution of the likelihood equations yields the vector $\underline{\hat{\theta}}$ of maximum likelihood estimates for $\underline{\theta}$. Appendix H provides background and techniques for optimization of expressions such as the likelihood function.

Example

We can illustrate maximum likelihood estimation by means of a Bernoulli distribution. Recall that the Bernoulli distribution specifies the probabilities of binary attributes, e.g., coin tosses (heads or tails), sex character (male or female), physiological condition (e.g., alive or dead), or capture status (captured or not captured). Suppose that we have a random sample of n bird nests, and $x = 1$ represents nest success (at least one bird fledges) and $x = 0$ represents nest failure (all eggs or nestlings are destroyed). Assume that 10 nests are observed, with the sample outcome $\underline{x} = \{1, 0, 1, 1, 1, 0, 0, 0, 1, 1\}$, or six successful and four failed nests. On assumption that the nest fates are independent and identically distributed, this is an outcome of a random sample of size 10 from a Bernoulli distribution, with unknown probability p of nest success. We can model the probability associated with y successes out of 10 nesting attempts with the binomial distribution:

$$f(y|p) = \binom{10}{y} p^y(1 - p)^{10-y}.$$

The likelihood function corresponding to six successes is therefore

$$L(p|y = 6) = \binom{10}{6} p^6(1 - p)^4.$$

The maximum of this likelihood can be obtained by taking the first derivative of the logarithm of the likelihood with respect to p, setting the result equal to 0, and solving for p. After some algebra this results in the equation

$$\frac{6}{p} - \frac{4}{1 - p} = 0,$$

or

$$\hat{p} = 0.6.$$

To show that this estimate maximizes the likelihood, we can construct the likelihood $L(p|y = 6)$ for the sample data and then substitute different values for p (Fig. 4.4). For any other value of p than 0.6 the likelihood will be seen to have a lower value than $L(0.6|y = 6)$.

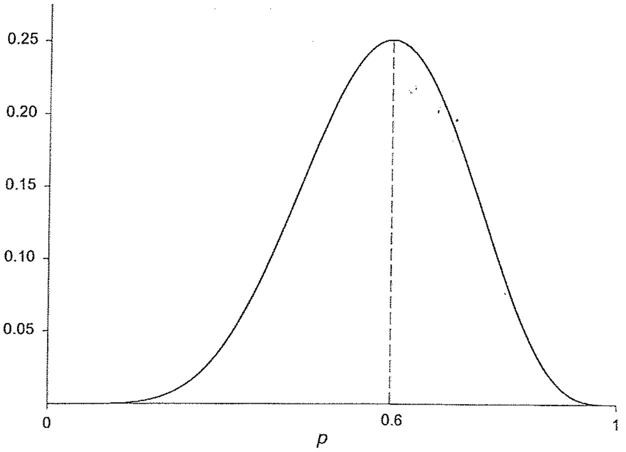

FIGURE 4.4 Example of maximum likelihood estimation with Bernoulli distribution, $p = 0.6$. Plots of the likelihood function $L(p|\underline{x}) = \binom{10}{6}p^6(1 - p)^4$ computed for sample data $\underline{x}' = (4, 6)$ in the text example, over a range of values for the parameter p. The maximum value of the likelihood function is attained by $\hat{p} = 0.6$, the maximum likelihood estimate for the data $\underline{x}$.

Example

The results from the example above can be generalized by allowing for an arbitrary number n of independent Bernoulli trials and any number y of successes. The appropriate statistical distribution for the situation is the binomial,

$$f(y|p) = \binom{n}{y} p^{y}(1 - p)^{n-y},$$

and the corresponding likelihood function for, say, $y = y_0$ successes is

$$L(p|y = y_0) = \binom{n}{y_0} p^{y_0}(1 - p)^{n-y_0}$$

As before, the maximum of this likelihood can be obtained by taking the first derivative of the logarithm of the likelihood with respect to p, setting the result equal to 0, and solving for p. This leads to the equation

$$y_0/p - (n - y_0)/(1 - p) = 0,$$

from which we get the intuitive maximum likelihood estimator

$$\hat{p} = y_0/n.$$

Example

A laboratory study involves testing the impacts on wildlife of the agrochemical diazinon. The objective of the study is to determine the mortality of passerine birds exposed to a particular concentration of diazinon. The investigation involves the use of American robins (*Turdus migratorius*) as a representative species, with an aim of estimating the probability p that a randomly chosen robin, when thus exposed, will die. The study design involves the exposure of 10 robins to diazinon, with the number of deaths recorded. The study is replicated three times, resulting in 3, 4, and 6 deaths. From probability theory we recognize that the appropriate distribution function for each replication in the experiment is the binomial B(10, p), where p is the probability of death. The replications are independent of each other by design, so the joint distribution for x_1, x_2, and x_3 is the product

$$f(x_1, x_2, x_3|p) = \binom{10}{x_1}\binom{10}{x_2}\binom{10}{x_3} p^{\Sigma_i x_i}(1 - p)^{30-\Sigma_i x_i},$$

with a likelihood function of

$$L(p|2, 3, 5) = \binom{10}{2}\binom{10}{3}\binom{10}{5} p^{10}(1 - p)^{20}$$

for this problem. Differentiation of the log likelihood yields the likelihood equation

$$10/p - 20/(1 - p) = 0,$$

which has the solution $\hat{p} = 0.33$. Based on data from the study and the binomial structure of the probability density function, the most likely value for mortality rate under the study conditions is 0.33. Note again that the estimate $\hat{p}$ depends on the particular values of x_1, x_2, and x_3. If the investigation were run again, different values for these random variables likely would be obtained, resulting in a different value for the estimate. It is in this sense that $\hat{p}$ can be thought of as a random variable, with its own distribution of values that is inherited from the distribution $f(x_1, x_2, x_3)$.

Maximizing a likelihood function involves the choosing of an estimate for each parameter in the probability distribution. For example, there are two parameters for the distribution of our quarterback's passing success for five randomly chosen passes. Maximizing the corresponding likelihood function involves the estimation of both parameters, based on the same data for passing success. Because the same data are used to estimate both parameters, the estimators typically are not statistically independent. As a general rule, the estimation of multiple parameters with the same data results in covariation among estimators.

Maximum likelihood estimation has some very strong advantages over other estimation techniques:

- The maximum likelihood estimator $\hat{\theta}$ has an approximately normal distribution for large sample sizes. Furthermore, its distribution converges asymptotically to a normal distribution as sample sizes increase (Appendix E).
- Though the estimator $\hat{\theta}$ may be biased, it is asymptotically unbiased in the sense that the expected value of $\hat{\theta}$ converges to the parameter θ as sample sizes increase.
- The variance of the estimator $\hat{\theta}$ is asymptotically minimum, in that $\hat{\theta}$ has the least variance of all unbiased estimators of θ when sample size is large.
- With some mathematical manipulation the variances and covariances of maximum likelihood estimators can be approximated directly from the likelihood function, using the "information matrix" (Appendix F).

It is because of these and other useful properties that we focus on maximum likelihood estimation in this book.

4.2.3. Confidence Intervals and Interval Estimation

Because an estimator of a population parameter is based on a random sample, the estimator is a random variable, with its own distribution. A key question is how good the estimator is, in the sense of being "close" to the parameter of interest. At issue is the relative confidence one has that an estimate obtained from random sampling accurately represents the parameter.

Approaches to this question constitute the subject matter of confidence intervals and interval estimation. Interval estimation is based on knowledge about the estimator's distribution, which can be obtained from the underlying sampling distribution. We focus here on the normal distribution, because maximum likelihood estimators (MLEs) are approximately normally distributed, asymptotically unbiased, with variances and covariances that can be derived directly from the likelihood function. Thus if $\hat{\theta}$ is an MLE with approximate variance $\widehat{\text{var}}(\hat{\theta})$, its distribution is approximately normal with distribution $N(\theta, \widehat{\text{var}}(\hat{\theta}))$. Furthermore, the frequency of occurrence of values for $\hat{\theta}$ in an interval (a, b) is simply the area under the corresponding normal curve between a and b.

A "confidence interval" for a parameter θ utilizes the distribution of $\hat{\theta}$. The idea is to identify an interval for values of $\hat{\theta}$ that is, say, 95% certain to include θ. Thus, to specify properly a confidence interval, both the interval length and the probability level associated with the interval are required. The procedure is first to identify the desired probability (the "confidence level") of including θ in the interval. Then the corresponding range of values for $\hat{\theta}$ is expressed in terms of θ. Finally, the expression is mathematically manipulated to identify confidence interval bounds on θ.

For example, one might seek a 95% confidence interval for θ based on the MLE $\hat{\theta}$. As indicated above, MLEs are approximately normally distributed estimators that are asymptotically unbiased. Therefore the MLE $\hat{\theta}$ is within 1.96 standard deviations of θ with probability 0.95, i.e.,

$$\theta - 1.96\sqrt{\widehat{\text{var}}(\hat{\theta})} < \hat{\theta} < \theta + 1.96\sqrt{\widehat{\text{var}}(\hat{\theta})}, \quad (4.4)$$

with probability 0.95, where both $\hat{\theta}$ and $\widehat{\text{var}}(\hat{\theta})$ are determined from the likelihood function. After some algebraic manipulation these inequalities can be rewritten as

$$\hat{\theta} - 1.96\sqrt{\widehat{\text{var}}(\hat{\theta})} < \theta < \hat{\theta} + 1.96\sqrt{\widehat{\text{var}}(\hat{\theta})}, \quad (4.5)$$

which exhibits a 95% confidence interval

$$\left(\hat{\theta} - 1.96\sqrt{\widehat{\text{var}}(\hat{\theta})},\ \hat{\theta} + 1.96\sqrt{\widehat{\text{var}}(\hat{\theta})}\right)$$

for the parameter θ. Note the distinction between expression (4.4), in which a fixed interval contains the random variable $\hat{\theta}$, and expression (4.5), in which a random interval contains the fixed parameter θ. Both expressions correspond to the same probability, so the interval in expression (4.5) is 95% certain to include θ. Said differently, replicated confidence intervals constructed as in expression (4.5) will include the parameter θ with 95% frequency. If one wishes to be even more certain that the interval includes θ, one can increase the interval length; i.e., the value 1.96 can be replaced by a larger number. The appropriate value for a given confidence level can be found in standard lookup tables in most statistics textbooks.

In the above example, confidence intervals are obtained by invoking the asymptotic properties of the MLEs, in particular the property that the estimates are asymptotically normally distributed. Small sample sizes and other factors can result in estimates that are not well represented by normal theory. Then confidence intervals based on a normal approximation may overrepresent the frequency with which the parameter is included in them. For some statistical models (e.g., the binomial) exact methods are available to compute confidence intervals in lieu of a reliance on normal theory. A more general methodology, applicable to all MLEs, is the method of profile likelihood. In profile likelihood, the likelihood function is used to calculate a confidence interval on a parameter θ_0 based on the function

$$\varphi(\theta_0) = 2 \ln \left[\frac{L(\hat{\theta}_0, \underline{\hat{\theta}})}{L(\theta_0, \underline{\hat{\theta}})}\right],$$

where $L(\hat{\theta}_0, \underline{\hat{\theta}})$ is the likelihood function evaluated at the MLEs for all parameters, and $L(\theta_0, \underline{\hat{\theta}})$ is the likelihood function evaluated at the MLEs of the other model parameters $\underline{\theta}$ and with θ_0 varying over its admissible range. It can be shown that the random variable $\varphi(\theta_0)$ is asymptotically distributed as chi-square with one degree of freedom (see Appendix E for a discussion of the chi-square distribution). A $(1 - \alpha)$ confidence interval on θ_0 may be obtained by solving for θ_0 in

$$\varphi(\theta_0) = \chi_1^2(\alpha), \quad (4.6)$$

where $\chi_1^2(\alpha)$ is the $(1 - \alpha)$ percentile of the chi-square distribution with one degree of freedom (Buckland *et al.*, 1993). There typically are two solutions to Eq. (4.6), and the confidence interval consists of all values θ_0 between them.

Example

Mule deer ($n = 100$) are outfitted with radio transmitters to estimate survival over winter (a 90-day study period). Of 100 deer, 10 animals die during the study period, no radio transmitters fail, and all the animals remain in the study area. The likelihood for the survival parameter p is therefore

$$L(p|10) = \binom{100}{10} p^{90}(1-p)^{10},$$

with a maximum likelihood estimate for p of $\hat{p} = 0.9$. The profile likelihood interval is obtained by

$$2 \ln\left[\frac{L(\hat{p}, \hat{p})}{L(p, \hat{p})}\right] = \chi_1^2(0.05)$$

or

$$2[-14.118 - \ln L(p, \hat{p})] = 3.841,$$

which has solutions for p at 0.788 and 0.966. Therefore the profile confidence interval (0.788, 0.966) is 95% certain to contain the parameter p. In contrast, an asymptotically normal confidence interval for p is formed by

$$\hat{p} \pm z(0.05)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

where $z(0.05) = 1.96$ is the 0.05 ordinate for a standard normal distribution [i.e., $P(z > 1.96) = 0.025$ if $z \sim N(0, 1)$; see Table 4.1]. This in turn yields the 95% confidence interval

$$(0.9 - (1.96)(0.03), 0.9 + (1.96)(0.03)) = (0.84, 0.96).$$

Thus the profile likelihood interval is a more conservative (i.e., wider) confidence interval than that produced by the normal approximation. As expected, the two intervals converge as sample size increases.

4.3. HYPOTHESIS TESTING

Closely associated with confidence interval estimation is the statistical testing of hypotheses, with an objective of determining whether parameters differ from hypothesized values. A testing procedure can be framed in terms of the comparison of null and alternative hypotheses. The null hypothesis

$$H_0\colon\ \theta = \theta_0$$

specifies some parameter value that is assumed prior to the test to be operative, but that is to be considered for possible rejection depending on the test results (see Chapter 2). The alternative hypothesis

$$H_a\colon\ \theta = \theta_a$$

specifies a second parameter value θ_a, to be considered as an alternative to θ_0 in the event the latter is rejected. As written, these hypotheses propose one of two values for the distribution parameter θ, with the true value of θ assumed to be either θ_0 or θ_a. The objective of testing is to determine whether, based on the data from a random sample, it is reasonable to conclude that the true value of θ is given by H_0. A straightforward method for testing H_0 comes directly from the procedure for confidence intervals: if θ_0 lies within the 95% confidence interval for θ, then H_0 is accepted (at the 5% significance level).

4.3.1. Type I and Type II Errors

The usual procedure for hypothesis testing requires specification of both the null and alternate hypotheses, as well as the specification of the significance level of the test. The test is essentially a binary decision process, in that the result is either to accept H_0 (over H_a) or reject H_0 (in favor of H_a). Thus there are two ways in which a testing procedure can reach a correct decision and two ways that it can make an error (Fig. 4.5). The correct decisions, of course, are to accept H_0 when it is true and to reject it when it is false. There are two types of incorrect decisions: rejecting H_0 when it is true (type I error) and accepting H_0 when it is false (type II error). The probabilities of making these two types of error are conventionally denoted as α and β, respectively.

One can guard against type I errors by increasing the significance level of the test. For instance, increasing the size of the confidence interval (i.e., by increasing the significance level of the test) makes it more likely that θ_0 will be included in the interval when H_0 is true. In general, decreasing the probability significance level of the test will decrease the probability of a type I error. However, larger confidence intervals also are more likely to include θ_0 even if H_0 is false. Thus, a decrease in the probability significance level *increases* the chances for a type II error at the same time that it

	REJECT	DO NOT REJECT
HYPOTHESIS TRUE	TYPE I ERROR	CORRECT DECISION
HYPOTHESIS FALSE	CORRECT DECISION	TYPE II ERROR

FIGURE 4.5 Possible outcomes for a simple hypothesis test. Off-diagonal entries correspond to correct inferences; diagonal entries correspond to incorrect inferences.

decreases the chances for a type I error. Clearly, there is a tradeoff between the two types of error that must be considered when establishing the significance level of a hypothesis test.

The testing of a simple null hypothesis (i.e., H_0 specifies a single value for θ_0) against a simple alternative hypothesis (i.e., H_a specifies a single value for θ_a) can be generalized to allow for composite null and alternative hypotheses, either or both of which can include a range of parameter values. For example, one might specify a null hypothesis that includes any value for θ that is less than some specified quantity, say θ^*, with the alternative hypothesis consisting of any parameter value larger than θ^*.

In particular, it is useful to consider simple null hypotheses against one-sided alternatives. As an example, one might investigate whether the body masses of males and females differ, on assumption that if there is a difference, it favors larger males. The null hypothesis for this situation specifies that the mean body mass of females is equal to that of males, with an alternative that the body mass of males is greater:

$$H_0:\ \mu_{\text{males}} = \mu_{\text{females}},$$

$$H_a:\ \mu_{\text{males}} > \mu_{\text{females}}.$$

Thus the test includes a one-sided alternative, with H_0 rejected only if the test results indicate that $\hat{\mu}_{\text{males}}$ is larger than $\hat{\mu}_{\text{females}}$ by some minimal amount. In general the testing procedure for simple versus one-sided hypotheses involves rejection of the null hypothesis if the test statistic exceeds some threshold value. The rejection region is associated with only one tail of the probability distribution of the test statistic.

In contrast, one could consider a simple null hypothesis against a two-sided alternative. Using the example involving body mass of males and females, the alternative model could allow the average body size of females to exceed that of males, and vice-versa:

$$H_0:\ \mu_{\text{males}} = \mu_{\text{females}},$$

$$H_a:\ \mu_{\text{males}} \neq \mu_{\text{females}}.$$

Rejection occurs if the mass for either sex exceeds that of the other by some critical amount, so that the rejection region is associated with both tails of the probability distribution of the test statistic.

The critical values of a test statistic signifying rejection will differ under one-sided versus two-sided alternatives, because the probability of type I error in the former case is allocated to one tail, whereas in the second case it is allocated to both tails (Fig. 4.6). This allows for a lower rejection threshold for the test statistic under a one-tailed test. It therefore is easier to recognize statistically significant differences with one-sided alternatives. This conclusion supports the intuitively appealing idea that biological structure is easier to recognize when its investigation is limited to a few, substantially different, alternatives.

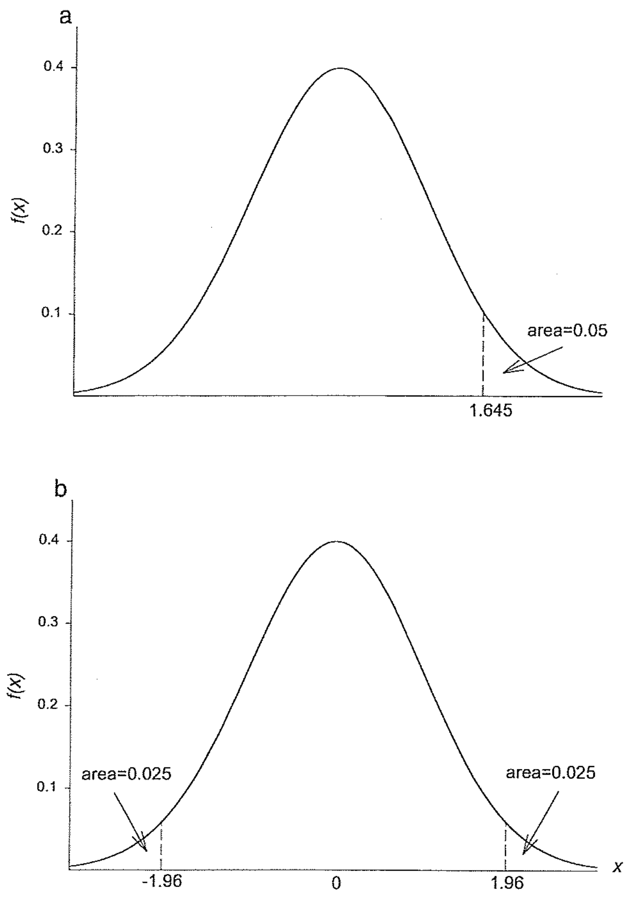

FIGURE 4.6 Hypothesis test with probability $\alpha = 0.05$ of a type I error in detecting a specified difference in population means. (a) One-sided alternative. (b) Two-sided alternative. Test statistic is distributed as $N(0,1)$.

4.3.2. Statistical Power

The ability of a test to reject false null hypotheses refers to the power of the test. Formally, test power is defined as $1 - P$ (type II error), so that powerful tests are unlikely to result in type II errors. The power of a test typically is high when (1) hypothesized parameter values in H_0 and H_a are quite different from each other, (2) the underlying sample distribution has low variance, and/or (3) the testing procedure is based on a sample of large size. Thus a strategy to protect against both types of errors is to set the probability significance level high (protecting against type I error) and to sample intensively enough to control against type II error. We will focus on sample sizes that maintain statistical power when we consider experimental design (Chapter 6).

4.3.3. Goodness-of-Fit Tests

Often it is important to determine the adequacy of a statistical model in characterizing field data. The idea is that if field data reflect an assumed statistical distribution it should be possible, by proper choice of distribution parameters, to demonstrate that the data "fit" the distribution. That is, it should be possible to choose parameter values such that the sample data conform to the statistical distribution underlying them. If the data reflect a distribution different from the one assumed, then there should be a "lack of fit," irrespective of parameter choice.

Goodness-of-fit procedures can be placed in the context of statistical testing, whereby the null hypothesis is that a particular model fits a set of field data and the alternative hypothesis is that the model does not fit the field data. However, there is an important difference between goodness-of-fit testing and the standard parametric testing procedures discussed earlier. Goodness-of-fit procedures use sample data to investigate the mathematical structure of a distribution, rather than specific values for its parameters. This is in contrast to the usual statistical procedures for parameter estimation and testing, which assume a known form of the distribution under investigation and utilize sample data to investigate specific values of the distribution parameters.

Goodness-of-fit testing is especially appropriate for multinomial distributions, for which there is a well-developed theory involving the use of maximum likelihood estimation. For problems for which the multinomial distribution is an appropriate statistical model the procedure is as follows:

1. The cell probabilities p_i, $i = 1, ..., k + 1$, of a multinomial distribution with $k + 1$ cells are modeled in terms of a vector $\underline{\theta}' = (\theta_1, ..., \theta_r)$ of parameters. The vector $\underline{\theta}$ typically contains fewer parameters than the number of cell probabilities. Cell probabilities for the model are designated as $p_i = p_i(\underline{\theta})$.

2. Maximum likelihood procedures are used to estimate the parameter $\underline{\theta}$. Let $\underline{\hat{\theta}}$ represent the MLE for $\underline{\theta}$.

3. The expected cell frequency $E(x_i) = np_i$ for each cell of the multinomial distribution is approximated with $n\hat{p}_i = np_i(\underline{\hat{\theta}})$, and the statistic

$$\chi^2 = \sum_{i=1}^{k+1} \frac{(n\hat{p}_i - x_i)^2}{n\hat{p}_i}$$

is calculated, where x_i is the observed cell frequency. This statistic is asymptotically distributed as a chi-square random variable with $k - r$ degrees of freedom, under the assumption that the mathematical form of the cell probabilities is correct.

4. The value χ^2 is compared to a tabulated chi-square value for the specified significance level of the test. A significance level of 0.95 often is chosen, so that a computed value of χ^2 exceeding the 0.95 chi-square quantile for $k - r$ degrees of freedom results in a rejection of the model. Model rejection essentially means that the multinomial model with cell probabilities parameterized by $p_i(\underline{\theta})$ is inadequate to characterize the data, i.e., the model is held not to "fit" the data.

Several alternatives to the above approach for assessing model fit exist and may have advantages in particular situations. The deviance, defined as

$$-2\ \ln[L(\underline{\hat{\theta}}|\underline{x})/L(\underline{\hat{\theta}}_{\text{saturated}}|\underline{x})],$$

describes the fit of a candidate model compared to a model containing as many parameters as independent observations, and is distributed approximately as chi-square with degrees of freedom equal to the difference between the candidate and saturated models (Agresti, 1990). The deviance is used in programs such as MARK (White and Burnham, 1999) to provide some indication of model fit. However, the chi-square approximation for either of these statistics is frequently poor in practice, especially when sample sizes are small, suggesting the need for approaches such as parametric bootstrapping (Appendix F) to more accurately assess model fit.

Example

The goodness-of-fit procedure can be illustrated with an example involving the banding of mallards in the preseason of year 1 and the subsequent recovery of bands during the next three hunting seasons. Assuming that survival and recovery rates are constant over time, the appropriate model for this situation is

$$\begin{aligned} f(\underline{x}|\underline{\theta}) &= \binom{n}{x_1,\ x_2,\ x_3} p_1^{x_1} p_2^{x_2} p_3^{x_3} \\ &\quad \times (1 - p_1 - p_2 - p_3)^{n - x_1 - x_2 - x_3} \\ &= \binom{n}{x_1,\ x_2,\ x_3} f^{x_1}(fS)^{x_2}(fS^2)^{x_3} \\ &\quad \times (1 - f - fS - fS^2)^{n - x_1 - x_2 - x_3}, \end{aligned}$$

where n is the number of birds banded and f and S are the recovery and survival rates, respectively. Four multinomial cells are defined for this model, with recoveries $(x_1, x_2, x_3, n - x_1 - x_2 - x_3)$ and expected recoveries $[np_1, np_2, np_3, n(1 - p_1 - p_2 - p_3)]$. Now assume that a total of 1000 mallards are banded in the preseason of year 1, and recoveries over the next three hunting seasons are $x_1 = 95$, $x_2 = 62$, and $x_3 = 39$. The corresponding likelihood function is

$$
\begin{aligned}
L(f, S|\underline{x}) &= \binom{1000}{95, 62, 39} p_1^{95} p_2^{62} p_3^{39} (1 - p_1 - p_2 - p_3)^{804} \\
&= \binom{1000}{95, 62, 39} f^{95} (fS)^{62} (fS^2)^{39} \\
&\quad \times (1 - f - fS - fS^2)^{804},
\end{aligned}
$$

from which the maximum likelihood estimates $\hat{f} = 0.095$ and $\hat{S} = 0.64$ are obtained. Thus the actual cell counts (95, 62, 39, 804) correspond to expected cell counts of (95, 61, 39, 805). Using these values in the goodness-of-fit statistic gives a value of $\chi^2 = 0.013$, with $k - r = 1$. When compared against the 0.05 significance value of 3.84 for a chi-square distribution with one degree of freedom, the model is seen to fit the data exceptionally well. Band recovery models of the kind used in this artificial example are developed in greater detail in Chapter 16.

Example

In an effort to determine the size of a population of fish, electroshocking sometimes is used in a removal experiment. A proposed model for this situation incorporates the assumptions that all fish have the same probability of removal and removal probability is constant over time. To test these assumptions for a particular species, 100 fish are subjected to an electroshocking experiment over four periods. The appropriate statistical model is

$$
\begin{aligned}
f(\underline{x}|\theta) &= \binom{100}{\underline{x}} p^{x_1} [(1 - p)p]^{x_2} \\
&\quad \times [(1 - p)^2 p]^{x_3} \\
&\quad \times [(1 - p)^3 p]^{x_4} \\
&\quad \times [(1 - p)^4]^{100 - \sum_{j=1}^{4} x_j},
\end{aligned}
$$

where p is the probability of removal of any randomly selected fish. The expected cell frequencies for this model are $E(x_j) = 100p(1 - p)^{j-1}$. Based on a sample with $(x_1, x_2, x_3, x_4) = (42, 23, 13, 10)$, the likelihood function is

$$
\begin{aligned}
f(\underline{x}|\theta) &= \binom{100}{\underline{x}} p^{42} [(1 - p)p]^{23} [(1 - p)^2 p]^{13} \\
&\quad \times [(1 - p)^3 p]^{10} [(1 - p)^4]^{12},
\end{aligned}
$$

with MLE $\hat{p} = 0.41$. Using $E(x_j) = 100(1 - p)^{j-1}$ the expected cell counts for the model are approximately (41, 24, 14, 9, 12). That the expected cell counts closely resemble the actual counts is confirmed by the goodness-of-fit statistic, which gives a value of $\chi^2 = 0.49$ with $k - r = 3$ degrees of freedom. When compared against a 0.95 significance level of 7.81 for a chi-square distribution with three degrees of freedom, the model is seen to fit the data. Removal studies of the kind used in this example are described in more detail in Chapter 14.

It is important to recognize the relationships between goodness of model fit and the probability significance of the goodness-of-fit test statistic. From the computing formula for step 3 above, it is clear that the goodness-of-fit statistic varies in magnitude to the extent that "expected" cell frequencies $n\hat{p}_i$ deviate from "observed" cell frequencies x_i. On assumption that the model is appropriate for the data, large deviations occur only infrequently, according to a chi-square distribution with $k - r$ degrees of freedom. Thus, the larger the computed value of χ^2, the smaller the probability that randomly collected data will generate a value that large or larger. The mathematical relationship between the magnitude of χ^2 and the probability significance level is parameterized by the chi-square "degrees of freedom" (see Appendix E).

Most of the statistical models we will use for estimation can be formulated in terms of cell counts for a multinomial distribution, and most have associated goodness-of-fit testing procedures. In particular, Part III presents goodness-of-fit tests for a class of statistical models that are useful for estimating population size and other population parameters.

4.3.4. Likelihood Ratio Tests for Model Comparisons

In the development of statistical models we are concerned not only with the adequacy of a model in characterizing data, but also with a comparison of the model with other models that differ in their parametric structures. Like goodness-of-fit testing, model comparison procedures can be seen as an example of a hypothesis test. The difference is that the general alternative of a goodness-of-fit test is replaced with an alternative of a specific model. Thus the test compares the fit of a hypothesized model versus the fit of an alternative model. Specifically, the null hypothesis is that the hypothesized model fits the data as well as the alternative model; the alternate hypothesis is that the alternative model fits the data better.

Typically the alternative model is more general in its parametric structure than the model of the null hypothesis, so that the null hypothesis can be couched in terms of restrictions on the parameter structure of the alternative model. The objective of a model comparison is to determine whether the fit of a hypothesized model can be improved by relaxation of its parameter restrictions. For example, the two-parameter model above for band recovery data can be generalized to include time-specific survival rates, and a comparison of the restricted and generalized models would give an indication of the importance of the additional parameters in characterizing the data.

As with goodness-of-fit testing, maximum likelihood estimation theory provides a mechanism for model comparisons. Let $\hat{\underline{\theta}}$ be the MLE corresponding to a model with parameterization $\underline{\theta}$ [e.g., $\underline{\theta}' = (S_1, ..., S_k, f)$], and $\hat{\underline{\theta}}_0$ be the MLE for a model with $\underline{\theta}$ restricted [e.g., $\underline{\theta}' = (S, f)$ with $S_1 = \cdots = S_k = S$]. Let $L(\hat{\underline{\theta}}|\underline{x})$ and $L(\hat{\underline{\theta}}_0|\underline{x})$ represent the likelihood function evaluated at $\hat{\underline{\theta}}$ and $\hat{\underline{\theta}}_0$, respectively. Then the likelihood ratio statistic

$$\chi^2 = -2 \ln[L(\hat{\underline{\theta}}_0|\underline{x})/L(\hat{\underline{\theta}}|\underline{x})] \tag{4.7}$$

is asymptotically distributed as a chi-square random variable, on condition that the restricted model is correct. The degrees of freedom λ for this statistic are given by the difference in the number of independent parameters in $\underline{\theta}$ and $\underline{\theta}_0$ (in our example $\lambda = k - 1$ degrees of freedom, because $\underline{\theta}$ and $\underline{\theta}_0$ contain $k + 1$ and two parameters, respectively).

A procedure for comparing models is as follows:

1. Goodness-of-fit procedures are used to identify a general statistical model "fitting" the data. Model generality is defined here in terms of the number of independent parameters, with the most general model defined by a lack of restrictions on the model parameters. Denote the corresponding parameter vector by $\underline{\theta}$ and its MLE by $\hat{\underline{\theta}}$. Let k_g denote the dimension of $\underline{\theta}$ and $L(\hat{\underline{\theta}}|\underline{x})$ represent the likelihood function evaluated at $\hat{\underline{\theta}}_0$.
2. Restrictions are imposed on the parameters in $\underline{\theta}$, to produce a restricted parameterization $\underline{\theta}_0$ for the model. The corresponding likelihood function $L(\underline{\theta}_0|\underline{x})$ is evaluated at the MLE $\hat{\underline{\theta}}_0$ of $\underline{\theta}_0$. Denote by k_r the dimension of the reduced parameter $\underline{\theta}_0$.
3. The likelihood ratio statistic χ^2 is calculated as above. This statistic is asymptotically distributed as a chi-square random variable with $k_g - k_r$ degrees of freedom, under the assumption that $\underline{\theta}_0$ is the appropriate parameterization.
4. The value χ^2 is compared to a tabulated chi-square value with $k_g - k_r$ degrees of freedom. If χ^2 exceeds the tabulated value for a specified significance level, the model with the more general parameterization is held to improve the fit of the model over that of the more restricted parameterization.

Example

In a study of harvest rates for mallards, an investigator releases 100 each of banded male and female birds prior to the hunting season, with the intent of examining harvest recoveries from each release sample. The objective is to determine if there are differences in harvest rates between the sexes. Let p_{m} and p_{f} represent the harvest probabilities for male and female birds respectively. Assuming that the harvest of males is independent of that of females, an appropriate statistical model is

$$f(x_{\mathrm{m}}, x_{\mathrm{f}}) = \binom{100}{x_{\mathrm{m}}} p_{\mathrm{m}}^{x_{\mathrm{m}}}(1 - p_{\mathrm{m}})^{100-x_{\mathrm{m}}} \binom{100}{x_{\mathrm{f}}} p_{\mathrm{f}}^{x_{\mathrm{f}}} \times (1 - p_{\mathrm{f}})^{100-x_{\mathrm{f}}}.$$

If $(x_{\mathrm{m}}, x_{\mathrm{f}}) = (30, 17)$ the likelihood function is

$$L(p_{\mathrm{m}}, p_{\mathrm{f}}|30, 17) = \binom{100}{30} p_{\mathrm{m}}^{30}(1 - p_{\mathrm{m}})^{70} \binom{100}{17} p_{\mathrm{f}}^{17} \times (1 - p_{\mathrm{f}})^{83},$$

and the maximum likelihood estimate for $\underline{p}$ is $\hat{\underline{p}} = \{0.30, 0.17\}$. Substituting these values into the likelihood function yields

$$L(\hat{p}_{\mathrm{m}}, \hat{p}_{\mathrm{f}}|30, 17) = \binom{100}{30} (0.30)^{30}(0.70)^{70} \times \binom{100}{17} (0.17)^{17}(0.83)^{83}.$$

Maximum likelihood estimates can be obtained in a similar way under the hypothesis of equal harvest rates, i.e., $p_{\mathrm{m}} = p_{\mathrm{f}} = p$. Under these conditions the likelihood function is

$$L(p|30, 17) = \binom{100}{30} p^{30}(1 - p)^{70} \binom{100}{17} p^{17}(1 - p)^{83},$$

with maximum likelihood estimate $\hat{p} = 0.235$. Substituting this value into the likelihood function yields

$$L(\hat{p}|30, 17) = \binom{100}{30} (0.235)^{30}(0.765)^{70} \times \binom{100}{17} (0.235)^{17}(0.765)^{83},$$

and Eq. (4.7) gives a likelihood ratio statistic of

$$\begin{aligned} \chi^2 &= -2 \ln[L(\hat{\underline{\theta}}_0|\underline{x})/L(\hat{\underline{\theta}}_{\mathrm{a}}|\underline{x})] \\ &= 4.749. \end{aligned}$$

Comparison of this value with a tabulated chi-square value of 3.84 for one degree of freedom and 5% significance indicates that the less restrictive parameterization significantly improves the fit of the model (at the 5% significance level). Thus, the study results provide evidence that there are differences in parameters for the two models, and in particular, that the harvest rates differ between sex categories.

The above sequence of steps began with an initial goodness-of-fit test for the more general model. In fact, the theory underlying likelihood ratio testing is based on the assumption that the more general model (corresponding to the alternate hypothesis) provides an adequate fit to the data. If the goodness-of-fit test for the most general model under consideration provides evidence of lack of fit, then the procedure for testing between models must be modified. Lack of fit often is a result of overdispersion of the data, in which case the goodness-of-fit statistic can be used to compute a "variance inflation factor," which in turn can be used to translate the likelihood ratio test statistic into a new test statistic distributed as F (see Section 17.1.8) (Lebreton *et al.*, 1992).

4.4. INFORMATION-THEORETIC APPROACHES

Though appropriate for many situations, the likelihood ratio testing procedure is not always satisfactory for model selection. First, there are philosophical problems with treating model selection as a hypothesis testing problem versus an estimation problem, particularly when the data have not been collected under an experimental design (e.g., Burnham and Anderson, 1992). Second, model comparisons based on likelihood ratio tests can only be used when the parameter space under one likelihood is a nested subset of that of a more general alternative. We will encounter instances later in which the parameter spaces of two competing models, fit to a common data set, will not be nested. To illustrate, suppose that data in the previous example are collected by both age and sex, so that a general parameterization would allow for both age and sex variation in harvest rates:

$$\underline{\theta}' = (p_{\mathrm{am}}, p_{\mathrm{ym}}, p_{\mathrm{af}}, p_{\mathrm{yf}}),$$

where the subscripts a and y now denote age-specific (adults and young) harvest rates within each sex. Two restrictions on this parameterization are

$$\underline{\theta}_1' = (p_{\mathrm{am}}, p_{\mathrm{ym}}, p_{\mathrm{f}}),$$

where

$$p_{\mathrm{af}} = p_{\mathrm{yf}} = p_{\mathrm{f}},$$

and

$$\underline{\theta}_2' = (p_{\mathrm{m}}, p_{\mathrm{af}}, p_{\mathrm{yf}})$$

with

$$p_{\mathrm{am}} = p_{\mathrm{ym}} = p_{\mathrm{m}}.$$

Though the parameterizations $\underline{\theta}_1$ and $\underline{\theta}_2$ are subsets of $\underline{\theta}$ and can be formed from $\underline{\theta}$ by imposing constraints on θ, neither $\underline{\theta}_1$ nor $\underline{\theta}_2$ can be formed by constraining the parameter space for the competing model. Attempts to apply Eq. (4.7) would result in a chi-square statistic with zero degrees of freedom, because each model has the same number of parameters.

An alternate approach based on information theory addresses the tradeoff between model fit (which favors more parameters) and estimator variance (which favors fewer parameters) in an optimization rather than hypothesis-testing framework. The approach is based on a statistic known as Akaike's information criterion (AIC) (Akaike, 1973), which utilizes the likelihood for each model via the term $-2\ \ln(L)$ and a penalty term for the number of parameters in the model:

$$\mathrm{AIC} = -2\ \ln(L) + 2q, \tag{4.8}$$

where L is the likelihood for a model under consideration and q is the number of parameters in the model. The idea is to select the model for which AIC is minimum. Although Eq. (4.8), especially the "penalty," seems somewhat arbitrary, AIC has a strong theoretical basis in information theory (Burnham and Anderson, 1998). In addition to expression (4.8), other forms for AIC also can be used; these incorporate a correction for small sample size ($\mathrm{AIC_c}$) and a "quasilikelihood" adjustment (QAIC) for extra binomial variation (Section 17.1.8) (see Burnham and Anderson, 1998). Model selection based on minimization of an appropriate one of these information measures will account for the bias–variance tradeoff in model parameterizations and if properly applied should result in the selection of a "best approximating model," i.e., the best data-based approximation to "full reality" (Burnham and Anderson, 1998).

Example

Consider the previous waterfowl harvest example, in which data are collected on both age and sex of the harvested animals, and 100 individuals of each age–sex stratum are banded and released prior to harvest. Let $\{p_{\mathrm{af}}, p_{\mathrm{am}}, p_{\mathrm{yf}}, p_{\mathrm{ym}}\}$ represent the probabilities of harvest for adults and young (subscripts a and y) of both sexes (subscripts m and f), with $\{x_{\mathrm{af}}, x_{\mathrm{am}}, x_{\mathrm{yf}}, x_{\mathrm{ym}}\}$ the numbers harvested in each category. Then an appropriate statistical model is

$$f(x_{\text{af}}, x_{\text{am}}, x_{\text{yf}}, x_{\text{ym}}) = \binom{100}{x_{\text{af}}} p_{\text{af}}^{x_{\text{af}}}(1 - p_{\text{af}})^{100-x_{\text{af}}}$$

$$\times \binom{100}{x_{\text{am}}} p_{\text{am}}^{x_{\text{am}}}(1 - p_{\text{am}})^{100-x_{\text{am}}}$$

$$\times \binom{100}{x_{\text{yf}}} p_{\text{yf}}^{x_{\text{yf}}}(1 - p_{\text{yf}})^{100-x_{\text{yf}}}$$

$$\times \binom{100}{x_{\text{ym}}} p_{\text{ym}}^{x_{\text{ym}}}(1 - p_{\text{ym}})^{100-x_{\text{ym}}}.$$

If $(x_{\text{af}}, x_{\text{am}}, x_{\text{yf}}, x_{\text{ym}}) = (35, 20, 47, 44)$ the likelihood function is

$$L(p_{\text{af}}, p_{\text{am}}, p_{\text{yf}}, p_{\text{ym}}|35, 20, 47, 44) = \binom{100}{35} p_{\text{af}}^{35}(1 - p_{\text{af}})^{65}$$

$$\times \binom{100}{20} p_{\text{am}}^{20}(1 - p_{\text{am}})^{80}$$

$$\times \binom{100}{47} p_{\text{yf}}^{47}(1 - p_{\text{yf}})^{53}$$

$$\times \binom{100}{44} p_{\text{ym}}^{44}(1 - p_{\text{ym}})^{56},$$

with a maximum likelihood estimate $\hat{\underline{\theta}} = \{0.35, 0.20, 0.47, 0.44\}$. These estimates result in a value for the log likelihood function of $\ln L = -9.8464$ and AIC value of

$$\begin{aligned} \text{AIC} &= -2 \ln L + 2q \\ &= -2(-9.8464) + 8 \\ &= 27.6928. \end{aligned}$$

On the other hand, the parameterization

$$\underline{\theta}_1' = (p_{\text{a}}, p_{\text{ym}}, p_{\text{yf}})$$

results in the likelihood function

$$L_1(p_{\text{a}}, p_{\text{yf}}, p_{\text{ym}}|35, 20, 47, 44) = \binom{100}{35} p_{\text{a}}^{35}(1 - p_{\text{a}})^{65}$$

$$\times \binom{100}{20} p_{\text{a}}^{20}(1 - p_{\text{a}})^{80}$$

$$\times \binom{100}{47} p_{\text{yf}}^{47}(1 - p_{\text{yf}})^{53}$$

$$\times \binom{100}{44} p_{\text{ym}}^{44}(1 - p_{\text{ym}})^{56},$$

with the maximum likelihood estimates

$$\hat{\underline{\theta}}_1' = (0.275, 0.47, 0.44)$$

and log likelihood of $\ln L_1 = -12.695265$. The corresponding value of AIC for this parameterization is

$$\begin{aligned} \text{AIC} &= -2(-12.695265) + 6 \\ &= 31.3905. \end{aligned}$$

By comparison, the parameterization

$$\underline{\theta}_2' = (p_{\text{af}}, p_{\text{am}}, p_{\text{y}}).$$

yields the likelihood function

$$L_2(p_{\text{af}}, p_{\text{am}}, p_{\text{y}}|35, 20, 47, 44) = \binom{100}{35} p_{\text{af}}^{35}(1 - p_{\text{af}})^{65}$$

$$\binom{100}{20} p_{\text{am}}^{20}(1 - p_{\text{am}})^{80}$$

$$\times \binom{100}{47} p_{\text{y}}^{47}(1 - p_{\text{y}})^{53}$$

$$\binom{100}{44} p_{\text{y}}^{44}(1 - p_{\text{y}})^{56},$$

with maximum likelihood estimate

$$\hat{\underline{\theta}}_2' = (0.35, 0.2, 0.455)$$

and log likelihood of $\ln L_2 = -9.93716$. The value for AIC in this case is

$$\begin{aligned} \text{AIC} &= -2(-9.93716) + 6 \\ &= 25.874. \end{aligned}$$

Based on AIC values of 27.6928, 31.3905, and 25.874 for the parameterizations $\underline{\theta}$, $\underline{\theta}_1$, and $\underline{\theta}_2$, respectively, we conclude that parameterization $\underline{\theta}_2$ provides the best variance–bias tradeoff among the three alternatives.

Some important points about AIC should be noted. First, AIC is appropriate only for comparison among models that all have been fit to a common set of sample data; comparison among AIC values from models fit to different data sets is meaningless. Second, a hypothesis testing framework for model selection may be preferable in those situations in which an experimental or quasiexperimental design provides a context for testing predictions based on theory, models, or both (see Chapters 2 and 3). Third, whereas AIC can be used to rank a number of competing, nonnested models, it does not always result in clear selection of a single model. Small differences in AIC can be expected to occur by chance and thus are indicative of virtually identical information content in the competing models. Burnham and Anderson (1998) advocate the computation of "model weights" based on the difference between each model's AIC value and that of the lowest ranked model. These weights, which are normalized to sum to 1 over all models considered, are roughly interpretable as the probability a given model is the best approximation to truth among the models considered. Closely ranked models with high weights (e.g., >0.25) should be retained for further consideration with other model selection criteria. Alternatively, the AIC weights may be used to compute weighted averages of parameter estimates across all the models considered. Buckland *et al.* (1997) and Burnham and Anderson (1998) recommend taking into account the (weighted) deviation of model-specific estimates from such a weighted average and inflating variance estimates accordingly. This approach accounts for the uncertainty in estimation induced by the process of model selection and seems preferable to the usual approach of reporting only sampling variances, conditional on an assumed true model. In Section 17.1.8 we discuss in greater detail the computation of model weights and corresponding variance components for capture–recapture models.

4.5. BAYESIAN EXTENSION OF LIKELIHOOD THEORY

There is yet another important application of likelihood theory in statistical estimation and hypothesis testing, which is especially useful for updating one's understanding of biological processes via predictive models (Section 3.3.2; see also Chapter 24). This application utilizes Eq. (4.1), along with a Bayesian definition of probability as "a measure of the degree of belief in an outcome" (Lee, 1992). On substitution of hypothesis H_i for event E_2, and $\underline{x}$, a sample outcome, for event E_1 in Eq. (4.1), one has

$$P(H_i|\underline{x}) = \frac{P(H_i)P(\underline{x}|H_i)}{P(\underline{x})}. \tag{4.9}$$

The value $P(H_i)$ in this expression is called the prior probability for hypothesis H_i, because it precedes the collection of the sample data $\underline{x}$, and $P(H_i|\underline{x})$ is called the posterior probability for H_i by reason of its being computed posterior to data collection (see Appendix A). Assuming that

$$\underline{H}' = \{H_1, ..., H_n\}$$

represents the set of all possible alternative hypotheses under investigation, by Bayes' Theorem we can rewrite $P(\underline{x})$ as

$$P(\underline{x}) = \sum_{m=1}^{n} P(\underline{x}|H_m)P(H_m)$$

and therefore Eq. (4.9) can be expressed as

$$P(H_i|\underline{x}) = P(H_i)P(\underline{x}|H_i)/\sum_{m=1}^{n} P(H_m)P(\underline{x}|H_m). \tag{4.10}$$

Equation (4.10) now gives us a procedure for evaluating the probability of any hypothesis in the hypothesis set based on the prior probabilities $P(H_i)$ and the likelihoods of each model given the sample data.

Example

Suppose there are three competing biological models (i.e., three hypotheses H_1, H_2, and H_3) for a system of interest, to which are assigned the prior probabilities $P(H_1)$, $P(H_2)$, and $P(H_3)$, respectively, based on previous information. Additional sample data $\underline{x}$ are collected in a field study, and these are used to obtain maximum likelihood estimates under each model. When evaluated at their respective maximum likelihood estimates, the three likelihood functions take values of $P(\underline{x}|H_1) = 0.10$, $P(\underline{x}|H_2) = 0.20$, $P(\underline{x}|H_3) = 0.15$. These values then can be used to compute the posterior probability of each hypothesis using Eq. (4.10). For example, equal prior probabilities result in

$$\begin{aligned} P(H_1|\underline{x}) &= P(H_1)P(\underline{x}|H_1)/\sum_{m=1}^{3} P(H_m)P(\underline{x}|H_m) \\ &= \frac{1/3(0.10)}{1/3(0.10) + 1/3(0.20) + (1/3)0.15} \\ &= 0.22. \end{aligned}$$

This same approach can be applied to determine $P(H_2|\underline{x}) = 0.44$ and $P(H_3|\underline{x}) = 0.33$. The new triple {0.22, 0.44, 0.33} of hypothesis probabilities reflects the fact that hypothesis H_2 fits the data $\underline{x}$ better than the other two hypotheses, and hypothesis H_1 fits the data more poorly than the others. These posterior probabilities now can serve as prior probabilities, to be updated with additional data in subsequent investigations.

In many situations the set $\underline{H}'$ of hypotheses consists of specific values θ_i that a parameter θ may assume. Then Eq. (4.10) can be written as

$$P(\theta = \theta_i|\underline{x}) = P(\theta_i)P(\underline{x}|\theta_i)/\sum_{m=1}^{n} P(\theta_m)P(\underline{x}|\theta_m).$$

If the prior probability distribution for θ is characterized by a continuous probability density function $f(\theta)$, the updating process is governed by

$$f(\theta|\underline{x}) = f(\theta)\, f(\underline{x}|\theta)/\int_\theta f(\nu)\, f(\underline{x}|\nu)\, d\nu.$$

In either case, once we have obtained updated (posterior) probabilities for the parameter values under consideration, these in turn can be used as new prior probabilities, in anticipation of another round of data collection and posterior updating (see Appendix A). This provides a powerful procedure for sequential updating, as we will see in Chapter 24 (also see Section 3.3.2).

4.6. DISCUSSION

In this chapter we have provided basic principles for estimation and hypothesis testing of parameters for animal populations. In Chapters 5 and 6 we explore how surveys and controlled experiments can be designed to assure that data are suitable for estimating model parameters and making comparisons of parameters over space or time, or with respect to other attributes. This background will be utilized in Part III to focus on statistical modeling techniques that are appropriate for animal populations, taking into account the conditions under which data from populations and communities are collected. These conditions affect the nature of inferences that can be made and make it necessary to develop specialized statistical models. In Part IV we explore the use of both deterministic and stochastic population models for optimal decision making in an adaptive framework, wherein monitoring and estimation contribute to the simultaneous pursuit of understanding and management of animal populations.