Calculation of Bayes Posterior Probability Distributions for Key Population Parameters

Carl Walters and Donald Ludwig

Fisheries Centre, University of British Columbia, Vancouver, BC V6T 1Z4, Canada

Walters, C., and D. Ludwig. 1994. Calculation of Bayes posterior probability distributions for key population parameters. Can. J. Fish. Aquat. Sci. 51: 713–722.

The Bayes posterior probability distribution is a powerful way to represent uncertainty in fisheries stock assessments, and can be calculated for key population and policy parameters of practically any population dynamics model. But the calculation is unwieldy when probabilities are to be assigned to a large grid of parameter combinations. The computational burden can be reduced substantially by analytically integrating over at least two "nuisance parameters" that occur in most assessment models: the observation error variance and the catchability coefficient. This simplification allows the analyst and manager to focus more easily on population parameters (stock size, slope of recruitment curve) that are of direct policy interest.

La distribution bayésienne de probabilité a posteriori constitue un outil puissant pour représenter les incertitudes dans les évaluations des stocks halieutiques, et peut être calculée pour des paramètres clés de la population et des politiques dans pratiquement n'importe quel modèle de la dynamique des populations. Toutefois, ce calcul est trop compliqué lorsqu'il faut attribuer des probabilités à une large grille de combinaisons de paramètres. Le fardeau du calcul peut être fortement réduit si on fait une intégration analytique d'au moins deux paramètres dérangeants qui sont présents dans la plupart des modèles d'évaluation : la variance de l'erreur d'observation et le coefficient de capturabilité. Cette simplification permet à l'analyste et au gestionnaire de s'intéresser plus directement aux paramètres sur les populations (taille du stock, pente de la courbe du recrutement) qui présentent un intérêt direct pour les politiques.

Received May 11, 1993
Accepted October 25, 1993
(JB921)

Reçu le 11 mai 1993
Accepté le 25 octobre 1993

Bayesian statistical procedures offer a way to provide information to fisheries managers about uncertainty in stock assessments. Assignments of probabilities to alternative parameter values and outcomes, rather than just best estimates and limits, is a fundamental part of recent approaches to risk assessment (Francis 1992; Restrepo et al. 1992; Hilborn et al. 1993) and adaptive management (Walters 1986; Sainsbury 1988; Hilborn and Walters 1992). By associating odds or probabilites with the many possible outcomes of each management decision option, analysts can provide decision-makers with more precise information for weighing choices about where to manage along the fundamental trade-off relationship between risk and opportunity associated with progressively more optimistic policy options. Probability calculations do not answer the social and economic question of where to operate along this trade-off, but do at least allow the choice to be made by objective and quantitative criteria.

The information is presented in the form of probability distributions for uncertain parameters such as maximum sustainable yield (MSY) and current stock size, rather than ranges or confidence limits on point estimates. Surprisingly, it often turns out to be no more difficult to calculate such probability distributions than to find single "best-fitting" point estimates and confidence limits. However, the calculation rapidly becomes intractable as the number of uncertain parameters is increased, since it must be done over a grid of parameter combinations and the number of such combinations increases geometrically with the number of parameters. This is particularly annoying when some of the parameters, such as measurement variances and catchability coefficients, represent only scaling effects in the data and are not of any direct policy interest (i.e., are "nuisance" parameters).

Here we offer suggestions for keeping the number of parameters small, and show that variance and catchability parameters can be eliminated entirely from the calculation by replacing the usual probability densities with a "marginal kernel" representing the integral of the densities over the nuisance parameters. Posterior probabilities calculated from this kernel for important population parameters represent the total probabilities for these parameters, accumulated over all possible values of the nuisance parameters. Consequently the distributions calculated from the population parameters provide a more accurate assessment of uncertainties than would be obtained by assuming particular values for the nuisance parameters.

The presentation proceeds in four steps. We first review the basic issue by presenting a general framework for Bayesian assessment of population dynamics parameters. Then we show how to simplify this framework by integrating over nuisance parameters. Next we provide simulation tests showing how the exact integration results compare with approximations that have been used in the past. Finally we provide warnings for practical users of the equations.

Overview: Bayesian Approach to Stock Assessment

Bayesian statistics begins with a fundamentally different definition and more intuitive notion of "probability" than

most scientists are taught in introductory statistics courses (Box and Tiao 1973; Berger 1985; Lee 1989; Press 1989). Classical statistics uses a very restricted definition of probability based on the expected frequency of occurrence of events under random sampling of possible events from well-defined sample spaces; parameters are viewed as fixed but unknown features of such spaces. To a classical statistician, it is meaningless to speak about the "probability of a parameter value"; any particular value either is or is not the true one, and probability statements are restricted to how likely it is to see various patterns in the data or in certain statistics for a fixed parameter value. One outcome of such a restrictive definition is the cumbersome notion of a confidence limit, which is described with phrases such as "this limit represents the largest (or smallest) parameter value for which there is a 5% or higher chance of obtaining the observed parameter estimate."

Bayesian calculations begin instead with a broader definition of "probability" as a measure of credibility or degree of belief. Under this definition, it is perfectly acceptable to talk about the probability of a particular parameter value being the correct one, provided we are careful about how this probability is calculated in relation to the data. The correct procedure is provided by Bayes' Theorem about joint and conditional probabilities of events in sample spaces. Classical statisticians do not dispute this theorem, only the manner in which it is employed by Bayesians. When applied to parameters and data, the theorem can be stated in words as follows:

$$\left(\begin{array}{l}\text{Probability of a}\\ \text{parameter value}\\ \text{given the data}\end{array}\right) = \frac{\left(\begin{array}{l}\text{Probability of the}\\ \text{data given the}\\ \text{parameter value}\end{array}\right) \times \left(\begin{array}{l}\text{Prior probability of}\\ \text{the parameter value}\end{array}\right)}{\text{Total probability of the data}}.$$

In applying this theorem, the data are regarded as fixed and the possible parameter values vary. When regarded as a function of the parameter values for fixed data, the "probability of the data given the parameter value" is called the likelihood function. The "total probability of the data" is the sum of the numerator over all parameter values. Division by the denominator can be viewed simply as a normalization to make the probabilities sum to unity; hence, this calculation can often be omitted or deferred until the last step of a series of calculations. The "probability of a parameter value given the data" is called the "Bayes posterior probability" for that value. Classical statisticians object to the specification of the "Prior probability of the parameter value", which represents a nonscientific (i.e., not based on the data) degree of credibility or judgemental weighting assigned to the parameter value. Obviously, one can generate misleading posterior probabilities by adjusting the prior probability to reflect personal beliefs and biases. Specification of the prior parameter distribution is a crucial aspect of the analysis, and it should never be done in a routine or merely conventional way. We address this issue in more detail below.

The importance of the distinction between Bayesian and classical definitions of probability is crucial when we try to answer fundamental management questions of the form "what are the odds that policy X is sustainable" for a range of possible policies (X's), given some precise definition of "sustainable". To answer such questions in the face of uncertainty about population parameters, we must examine a range of parameter values, since the total probability of sustainability is a cumulative probability summed (or integrated) over all possible parameter values. The contribution to the total probability of sustainability for any one parameter value in this sum is the probability that the value is the true one given the data, times the probability of sustainability if the parameter value is the true one. For complex models the sum may be estimated most efficiently by using Monte Carlo simulation methods (Restrepo et al. 1992). But under the classical definition of probability, we cannot even form such an obvious sum because the "probability that the value is the true one given the data" does not exist! One notion for supposedly avoiding this difficulty is the "confidence distribution" used, for example, by Francis (1992), generated by calculating the probability of obtaining the observed parameter estimates for a small neighborhood around each possible parameter value; but examination of how such distributions are calculated and renormalized reveals that they are just Bayes posterior distributions computed by a numerically cumbersome route involving sampling distributions of parameter estimates rather than direct inference from the data.

Sometimes it is appropriate to initially weight all parameter values equally, i.e., choose a uniform prior. When using a uniform prior, the Bayes posterior probability calculation reduces to the remarkably simple formula

$$\left(\begin{array}{l}\text{Probability of}\\ \text{a parameter value}\\ \text{given the data}\end{array}\right) = \frac{\left(\begin{array}{l}\text{Likelihood of the}\\ \text{parameter value}\\ \text{given the data}\end{array}\right)}{\left(\begin{array}{l}\text{Sum of likelihoods over}\\ \text{all parameter values admitted}\end{array}\right)}.$$

Compare this approach with maximum likelihood parameter estimation: in point estimation, one must be able to calculate the probability of observing the data given any parameter value, and one must then devise a formula or search algorithm for finding the parameter value that gives the highest likelihood (that value is called the "maximum likelihood estimate" of the parameter). For most fisheries assessments, the search is done by means of a nonlinear estimation algorithm which calculates the likelihood at many different parameter values along the way.

A very important point about the Bayesian assessment of posterior probabilities is that we go directly from the data to the posterior probability using the above formula, systematically assigning probabilities to a wide range of parameter values, without ever worrying about calculating point estimates along the way (except for some nuisance parameters as discussed below). In other words, we never worry about the sampling distribution of point estimators, which the classical statistician requires in order to provide confidence limits, nor do we pretend that one can define or compute point estimates with particular properties such as being unbiased or asymptotically efficient or minimum variance or minimum risk or whatever. The posterior probability distribution is used directly to obtain estimates of the probabilities of various outcomes of interest to managers, such as stock collapse or capacity to sustain higher catches.

Notice also that the posterior probability does not depend

on any constant multiplier terms that may appear in the prior or the likelihood function, and that are independent of the data and parameter values over which the sum (or integral) is taken. Below we will refer to a likelihood multiplied by the prior and stripped of such constants as the "posterior kernel". Usually this kernel is a quite simple equation involving familiar measures of goodness of fit, such as sums of squared deviations between observations and predictions.

Setting up a Bayesian estimation scheme involves three steps: (1) choice of a prior probability distribution for the parameters, (2) formulation of a likelihood function for the data used in the assessment, and (3) choice of a grid of parameter values over which the product of likelihood and prior is calculated. The following subsections discuss the first two of these steps. It is generally a simple programming or spreadsheet exercise to apply the formulae (calculate likelihood times prior) resulting from steps (1) and (2) for each parameter combination on a grid, and it is then trivial to rescale each likelihood times prior by the sum of all likelihoods times priors over the grid in order to provide the posterior probability for the parameter value at that grid point.

Selection of Prior Distribution

There has been much discussion about the best choice of prior probability distributions for Bayesian calculations. The discussion has centered on two quite distinct questions: (1) the best functional form to assume and (2) whether or not to use an "informative" prior based on expert judgement and past experience. The functional form question is mainly concerned with finding so-called "natural conjugate priors" that are analytically or computationally convenient to use; this question is not of much interest in the present context because most fisheries models make predictions that are highly nonlinear functions of the model parameters and hence always require a computationally intensive approach. Thus the main issue is whether to use an informative prior. Here we offer a few options and suggestions about potential pitfalls.

The simplest computational option is to assume a uniform prior over some specified, "reasonable" range of parameter values. This approach focusses debate on the relatively simple matter of deciding on reasonable limits, and amounts to assuming an uninformative prior provided the limits are wide enough to include most of the tails of the likelihood function. Even this simple approach involves at least two subtleties. First, if one chooses the prior to be zero outside a particular interval, say $q_1 \leq q \leq q_2$, then in effect one refuses to consider the possibility that q might lie outside this interval. If this interval is inappropriately chosen, the posterior distribution may be cut off at a point where it is substantially different from zero. The resulting inferences will be suspect in such a case. This possibility can be avoided by not making any cutoff, i.e., choosing an infinite interval on which the prior is constant. The posterior may then be restricted to an interval where the likelihood is sensibly different from zero. Second, choice of a uniform prior has different consequences, depending on the scaling or choice of variable. For instance, a catchability coefficient q cannot be negative. Therefore, a uniform prior would be nonzero on the interval $0 \leq q < \infty$ or a subset thereof. As we shall see later, sometimes $\log q$ is a more convenient variable to use than q itself. If the density of the random variable W is uniform, then

$$\Pr[w \leq W \leq w + dw] = dw.$$

But if $W = \log Q$ and $w = \log q$, then $dw = dq/q$. Hence

$$\Pr[q \leq Q \leq q + dq] = dw = dq/q.$$

Thus, if $w = \log q$ is uniform, then the density of q is dq/q. Similarly, if the density of q is uniform, the density of w is $e^w dw$. There are arguments in favor of either choice. Which is correct? We are unable to resolve this issue: a prudent procedure is to try both priors and determine whether there are substantial differences in the final inferences. If there are, this is an indication that the data are not very informative about the parameter in question. This fact should be reported as part of the analysis, so that decision-makers are not misled about the degree of risk that may be associated with one action versus another. In spite of these difficulties, we believe that an uninformative prior should be the default choice, since it eliminates the possibility that the analysis will be colored or distorted by the subjective beliefs that stock assessment scientists might carry.

A second option is to use an informative prior based on empirical distributions of parameter values reported for similar stocks. For some per-capita rate parameters that represent basic life history characteristics (such as natural mortality rate (M) and growth parameters), such distributions can be constructed from syntheses of parameter estimates in the literature, and it makes good sense to use such experience in the analysis. However, for parameters that represent abundance scaling and habitat size (carrying capacity K of logistic, unfished or natural biomass B_0), such empirical distributions are likely to be very broad and are potentially misleading (e.g., if literature values have been reported mainly for large, valuable stocks and you happen to be working on a small one).

The third option is to use an informative prior based not on empirical experience with other stocks but rather on arguments from basic biology (fecundity, longevity, maximum possible density, etc.) combined with scientific "intuition". We feel strongly that such subjective priors should be avoided whenever possible, for two reasons: (1) the track record of biologists at making intuitive predictions about ecological parameters has not been good, and we regularly are overconfident about what we "know"; and (2) it is very easy for past data and analyses of the same stock to creep into the thinking and judgement of scientists, especially about parameters that have been much debated in the past and about which some scientific concensus has been reached. Including past results of analysis in the prior causes a very dangerous type of circular reasoning or double accounting with the data.

Whatever option is chosen, care must be taken about hidden assumptions that can enter the analysis though choices of scaling for the parameters. Particularly for parameters such as the catchability coefficient that can physically take only positive values, it may be tempting to assume a uniform or flat prior distribution for the logarithm of the parameter. But this choice results in assignment of higher probabilities to larger stock sizes; below we provide an example of how large an effect such a choice can have on the assessment.

Calculation of the Likelihood Function

For most fisheries assessments, we deal with heterogeneous collection of independent observations $Y_1, \ldots, Y_n$ of quantities such as catch per unit effort (CPUE), mean fish

sizes, survey biomass estimates, etc. The simplest assumption is that each observation Y_i is a random variable sampled from a normal distribution with mean U_i and variance V_i, where U_i is some function of unknown population parameters. In this case, the likelihood part of the kernel needed for the posterior distribution calculation if the V_i are assumed known is just

$$(1) \qquad L = \exp\left(-\sum \frac{(Y_i - U_i)^2}{2V_i} \right).$$

The other factors in the normal probability density cancel out when calculating the posterior probability, since the posterior is the product of L times prior for a particular parameter value divided by the sum of such products over all such values. Thus the likelihood calculation is very simple under the assumption of normal (or log-normal) measurement errors, provided there is a clear specification of how to calculate the means U_i from population parameters of interest. Here is where stock assessment population models enter the analysis. Population models are used to predict the U_i in two steps: (1) set population model parameters at particular values (i.e., a grid point) for which a likelihood is to be calculated and (2) simulate the model over time for the period for which there are data, calculating predicted values for the U_i along the way. For example, suppose Y_i is the CPUE in year 8, and one is willing (foolishly!) to assume that $Y_i = qB_8$ where q is a catchability parameter and B_8 is the stock biomass in year 8. To predict B_8, one needs (1) population model parameters for initial stock size and growth/mortality/recruitment and (2) historical removals (catches) for years 1–7. For a particular combination of values of these parameters and a particular value of q, U_i is then simply q times the model prediction of B_8.

In the presentation below, we will assume unless stated otherwise that $Y_i = qB_i$ where i is an observation counter, B_i is any predicted state variable (biomass, number, mean size, etc.), and q is an unknown scaling or proportionality parameter. Under this assumption, each element of the sum of squares is of the form $(Y_i - qB_i)^2/V_i$.

It may not be obvious just how general is this framework for calculating likelihoods. One may use literally any model that is complex enough to predict the means of the observations. When only relative abundance data (e.g., CPUE) are available, the intuitive choice would be to use simple surplus production or delay-difference models. But we can as well do the predictions with a complex age-structured model, calculating overall biomass or catch rate predictions as the simulation proceeds (e.g., Francis 1992). In fact, the model need not even do its predictions forward in time. In virtual population (VPA) or cohort analysis models and in stock reduction analysis, we specify a terminal or current stock size (usually by specifying a terminal fishing rate parameter) and then work backwards in time to reconstruct historical stock sizes by adding natural mortalities and age-specific catches back into the simulated population.

Adaptive or tuning procedures in VPA (such as ADAPT; Gavaris 1988) refer to estimation procedures where the population model is simulated backward in time, with different current (starting) stock sizes (or equivalently, terminal fishing rates), attempting to find that parameter value (current stock size) for which the backward simulation provides a best fit to some relative abundance time series. Often the tuning procedure is used to simultaneously estimate some "nuisance parameters" (partial recruitment factors or relative vulnerabilities at age) as well. Viewed in this way, the key difference between use of VPA and fitting forward simulations to relative abundance data is in the use of catch-at-age data to avoid making assumptions about the stock–recruitment relationship (some assumption is unavoidable in forward simulations whereas individual recruitments are estimated directly with VPA). However, note that in either case, a crucial assumption is that the relative abundance index is proportional to stock size. When nuisance parameters and other input data to VPA are uncertain, assessment of posterior distributions for key policy parameters might be done more efficiently using Monte Carlo procedures as described in Restrepo et al. (1992) than using the grid-based procedures discussed below. Still, care must be taken in the use of such simulation procedures; Restrepo et al. actually did not compute the posterior distribution, but instead proposed to approximate it by the sampling distribution of parameter estimates evaluated at the best-fitting parameter values (they treated the best-fitting parameter values as correct when they used these to set the expected values of sampling distributions for VPA input values). This approximation is only asymptotically valid for much longer time series than are usually available for VPA tuning; it is avoided in the confidence distribution method of Francis (1992) by computing the sampling distributions for a range of parameter values.

For forward simulations in time, in our experience it is usually worthwhile to go beyond a simple surplus production model to at least the Deriso/Schnute delay-difference biomass/numbers dynamic model (Deriso 1980; Schnute 1985), unless the data are very uninformative (Ludwig and Walters 1985). This model allows the representation of important delay effects in recruitment (especially critical in modelling population changes early in fishery development for long-lived species). It also makes predictions about changes in mean body sizes, and these changes are an important signal for detecting changes in mortality rate due to fishing (and hence estimating total population size) even if age composition data are not available or are suspect for various reasions (Hilborn and Walters 1992).

An important point is that it is impossible to say a priori what the best model structure will be. One can obviously obtain misleading results from models that are too simple, but detailed and realistic models can also fail badly (in estimation of the best fishing policy) if they require estimation of too many parameters. Luckily, it is easy to test a variety of models once one has a computational framework for the parameter grid and likelihood calculation. Thus the best strategy is just to test a sequence of increasingly realistic models, and see what each has to say about the data. It is also possible to combine both simple and complex models, as suggested by Fournier and Warburton (1989).

Choosing Parameters to Represent in the Posterior Distribution

Once we step beyond surplus production models, it is necessary to deal with the proliferation of parameters associated with population age structure (recruitment parameters, natural mortality rate, growth curves, mesh selection). The following three conventions are gaining wide use (R. Hilborn and A. Punt, University of Washington, Seattle,

Wash., personal communication), and appear to be working very well.

First, it is usually necessary to treat the natural mortality rate M and growth curve parameters as known in advance, as in assessment methods like VPA. Some information about these parameters is contained in time series data on age composition and body size distribution moments of the catch, but this information is generally very weak unless the stock has been subject to strongly contrasting harvest regimes over time.

Second, it appears best to always include the initial, unfished biomass B_0 as a key parameter (or the current biomass for VPA models). There are at least three reasons for this: (1) B_0 is needed to start each simulation leading to a likelihood value, (2) B_0 is often a good, simple predictor of MSY (0.5 MB_0 type calculation), and (3) for forward simulations over time, B_0 automatically defines one recruitment parameter (the average recruitment rate at the unfished biomass level, equal to MN_0 where $N_0 = B_0/\bar{w}$ and $\bar{w}$ is average fish size in the absence of fishing). In surplus production models, B_0 is the carrying capacity (K) parameter anyway.

Third, consider including one parameter for the stock–recruitment curve slope in the Bayesian parameter grid; this slope basically measures how low the stock size can be driven before recruitment overfishing becomes evident, and hence is a key parameter in policy evaluation. For long-lived species, there will be no information about this slope parameter in the data until the fishery has been operating for at least as long as the age at recruitment. In surplus production models, the recruitment slope corresponds to the parameter for intrinsic rate of population increase (r) at low stock sizes. Notice that steps (2) and (3) result in two stock–recruitment parameters. One can then represent patterns of variation in recruitment rate between the extremes represented by the recruitment rate at unfished equilibrium and the slope at very low stock size by any functional form that seems appropriate, e.g., Beverton–Holt or Ricker.

Notice that these three steps result in a Bayesian estimation scheme that actually deals with three parameters corresponding to the K, r, and q of classical surplus production models. One difference from classical estimation methods is that it is possible to treat q as a nuisance parameter to be estimated as a conditional maximum likelihood value for each combination of K and r (B_0 and recruitment curve slope) included in the Bayes posterior distribution, or eliminate it entirely as shown in the next section.

For age-structured models, it is usually necessary to consider including further parameters such as those needed to characterize a mesh selection ogive. One may also want to look at recruitment models with more than two parameters. However, before proceeding to throw more parameters into the analysis, consider carefully the following points about parameter grids. Suppose one wants to generate a Bayes posterior distribution for p parameters. In our experience, it is necessary to represent each parameter at 40 or more discrete levels between some minimum and maximum values beyond which the posterior probabilities are vanishingly small. To get the posterior distribution, this means that it is necessary to do the likelihood calculation (run model and compute L) at 40^p combinations of parameter levels. If $p = 2$, this is not difficult with personal computers; 40^2 is just 1600 simulations, which for typical population models and 30 yr of data, an 80486-based computer can run through in about 10 s. But with a third parameter, the grid requires $1600 \times 40 = 64\,000$ simulations, which will require approximately 5–10 min. With a fourth parameter, expect to wait several hours for an answer, and it might be very difficult to visualize (or explain to a manager) a four-dimensional probability distribution. An option of course is to perform the large calculation and then add up the posterior probabilities over two or three of the parameters to provide a marginal distribution for one or two others that are of particular management interest, but in our experience this is usually not worth the trouble. Following the prescription of this section, including only B_0 and recruitment curve slope in the parameter set will generally cover about 90% of the management concerns and key uncertainties about long-term production potential.

One difficult problem is that the high probability parameter combinations often lie very close to a curved line of B_0 and r combinations (in classical terms, the joint confidence region for B_0 and r is a long, very narrow ellipse or banana-shaped region), especially if the variances V_i for the observations are relatively low or there are many years of data. What can happen in this case is that only a few scattered (or even no) grid combinations of B_0 and r happen to hit along the high probability ridge, so the calculated probability surface will look like a few peaks scattered along the line. None of these calculated peaks may even be close to the most probable parameter combination. To check for this problem, a simple trick is to artificially inflate the variances V_i to spread out the probability distribution so that the region of high probability combinations is more visible. To further ward against the problem, options are to (1) increase the number of grid points, (2) vary the parameter ranges over the grid, to "window in" on parts of the surface, and/or (3) set up a simple line search (one-dimensional nonlinear estimation) along one of the parameter dimensions. That is, for each B_0 on the grid, search for the most probable r value and plot the resulting line of high probability B_0–r combinations. This may actually take less computer time than calculating the probabilities of all r levels for each B_0.

Estimating Probability Distributions for Policy Parameters

The grid of posterior probabilities for population parameters can be used to estimate posterior probabilities for policy parameters that are derived from the population parameters. The basic procedure is to set up a grid of discrete, narrow ranges for each policy parameter and then find the total probability that the parameter lies within each of these ranges.

Suppose for example that a distribution is wanted for MSY = $rB_0/4$ predicted from a simple logistic surplus production model. To obtain this distribution, set up a series of MSY "bins", each representing a narrow range of MSY values, and set up an array to store the probability for each bin. Initially set these probability array values to zero. Then for each point in the (B_0,r) grid, calculate MSY and determine which MSY bin this falls in, and add the posterior probability assigned to (B_0,r) to the total probability for the bin. After doing this for all (B_0,r), each MSY bin should contain the total probability that MSY lies in the range represented by the bin.

One very interesting result of such calculations is that policy parameters are often much less uncertain than the

original population parameters. This happens when a wide range of combinations of population parameters result in the same value of the policy parameter. In the above example, MSY = $rB_0/4$, the range is easy to see: all r–B_0 combinations such that rB_0 is the same will predict the same value of MSY. Intuitively, this just says that the same MSY can be produced either by a small, productive stock or a large, unproductive one. Early assessment methods such as Gulland's (1961) plot of catch or CPUE versus effort relied heavily on this invariance property to avoid having to estimate population parameters.

Posterior Kernels Integrated over Parameters

We show now how the product of the prior and the likelihood kernel (1) introduced above can be replaced by an alternative expression that does not depend on the nuisance parameters V and q, representing the integral or marginal probability over these parameters. We note first that the likelihood part of the kernel can be written as a product of the form

$$L = L_1 L_2 \ldots L_m \tag{2}$$

where each of the m factors represents a collection of observations with homogeneous variance and constant scale parameter q. The prior density will be assumed to be independent of the parameter q. This assumption can be generalized to the case where the prior is proportional to power of q. Our notation will not display other parameters that may appear in the posterior. If the kth such data type has n observations, with variance V and scale parameter q and if V is assumed to be known, then

$$L_k = \exp(-\mathrm{SS}_k/2) \tag{3}$$

where

$$\mathrm{SS}_k = \frac{1}{V}[(Y_1 - qB_1)^2 + (Y_2 - qB_2)^2 + \ldots + (Y_n - qB_n)^2]. \tag{4}$$

Here, B_i is the model prediction of the unobserved state variable (biomass, mean body size, number-at-age, etc.) to which observation Y_i is assumed to be proportional (via the constant q). In order to reduce notational clutter, we shall consider only a single term L_k and drop the subscript k. We can conveniently write SS in terms of computational sums of squares as

$$\mathrm{SS} = \frac{1}{V}\left(\sum Y_i^2 - 2q\sum Y_i B_i + q^2 \sum B_i^2\right). \tag{5}$$

Minimization of SS with respect to q for a fixed set of population parameters leading to the predicted B_i leads to the conditional maximum likelihood estimator q^*:

$$q^* = \frac{\sum Y_i B_i}{\sum B_i^2} \tag{6}$$

and substituting (6) into (5) leads to the familiar decomposition

$$\mathrm{SS} = \frac{1}{V}[(n-1)s^2 + I(q - q^*)^2] \tag{7}$$

where

$$s^2 = \frac{\sum (Y_i - q^* B_i)^2}{n-1} \quad \text{and} \quad I = \sum B_i^2. \tag{8}$$

The quantity $(n-1)s^2$ is the sum of squared residuals and I is proportional to the information. The residual sum of squares is a function of the data and remaining parameters only, and hence contains no information about the dependence of the posterior on q. However, the magnitude of I determines the variance of the distribution of q.

One approximate Bayesian procedure is to replace q by q^* in (6), with the posterior kernel L^* replacing L:

$$L^* = \exp\left(\frac{-(n-1)s^2}{2V}\right). \tag{9}$$

This procedure ignores uncertainty about q entirely, and therefore it is risky when n is small (<20).

Such simplifications are not necessary if we assume a uniform prior probability distribution for q (all values equally likely a priori). In that case, the Appendix shows that the marginal density L^q (L integrated over q) is given by

$$L^q = \sqrt{\frac{V}{I}} L^*. \tag{10}$$

Here we use the superscript q to indicate the integration of L over q. Equation (10) can be used to eliminate the parameter q from Bayesian calculations where V is assumed known in advance, simply by substituting it for the corresponding L_k term in (2).

Elimination of the variance V from the calculation is more troublesome (e.g., see Hoenig et al. 1994). Jeffreys (1961) sugggested using prior probabilities for V proportional to $1/V$. The corresponding posterior is obtained by multiplying L^q by $1/V$ and restoring terms involving V that are not needed when L^q alone is used for numerical calculation, to form a likelihood × prior kernel. The resulting density must be integrated over V to obtain the marginal density:

$$L^{qV} = \int_0^\infty \frac{L^q}{V^{(n-1)/2}} dV. \tag{11}$$

The superscript qV denotes integration over q and V. It is shown in the Appendix that the result of this integration is simply

$$L^{qV} = (s^2)^{-m} \tag{12}$$

where $m = (n-1)/2$. Note that this kernel is very similar to (9), but with the exponential function of s^2 replaced by a power function. Writing the power function as $\exp[-m \log(s^2)]$ hints that (9) and (12) will give similar shapes of posterior distributions for the population parameters that determine s^2, as sample size n increases. We show by example in the next section that this is indeed the case.

Posterior kernels similar to (10) and (12) can also be substituted for posterior components L_k that contain direct sums of squares differences between data and model predictions, without unknown scaling factors q. For these cases, simply drop the $1/\sqrt{\sum B^2}$ term from the kernel in (10) (i.e., integrate L_k times a $1/\sqrt{V}$ prior kernel, over V only; the sum of squared differences between observed and predicted values then replaces s^2).

It is often best to assume that the measurement errors are multiplicative, and that Y has a log-normal distribution (e.g., $Y = qBe^v$, where v is normal with mean 0). In this case, if we replace Y_i with $Z_i = \log(Y_i/B_i)$ and q with $q' = \log q$, then

$$L_k = \frac{1}{V^{n/2}} \exp\left(-\frac{1}{2V}\sum (Z_i - q')^2\right). \tag{13}$$

In analogy with the preceding treatment, we can rewrite (13) for convenience in integrating over q' as

$$(14) \quad L_k = V^{-n/2} \exp\left(-\frac{1}{2V}[(n-1)s'^2 + n(q' - \hat{q}')^2] \right).$$

Here

$$(15) \quad \hat{q}' = \frac{\sum Z_i}{n}$$

and

$$(16) \quad s'^2 = \frac{\sum (Z_i - \hat{q}')^2}{n-1}.$$

Now if we assume a uniform prior for q', integration over q' results in the posterior kernel

$$(17) \quad L^{q'} = V^{-n/2} \exp\left(\frac{-(n-1)s'^2}{2V} \right)$$

within a nonessential factor. Integrating over V then results in the kernel

$$(18) \quad L^{q'V} = (s'^2)^{-m}$$

where $m = (n - 1)/2$ as before. Note that this is essentially the same kernel as (12), but with a simple change in the way the prediction errors are calculated.

It is also possible in principle to use equations similar to (9)–(18) to deal with a more general power model for the observations:

$$(19) \quad Y = q_1 B^{q_2}$$

where q_1 is a scaling parameter and the power parameter q_2 can represent both saturating ($q_2 < 1$) or accelerating ($q_2 > 1$) observation processes. This model has been suggested for example as a way to deal with saturating relationships between CPUE and stock size (Cooke and Beddington 1985).

The preceding logarithmic transformations work in this case as well: the nuisance parameters q_1, q_2, and V may be integrated out to yield a marginal kernel

$$(20) \quad L^{q_1 q_2 V} = (s''^2)^{-m}$$

where

$$(21) \quad s''^2 = \frac{1}{n-2} \sum_i [Z_i - (\hat{\theta}_1 + \hat{\theta}_2 W_i)]^2$$

$$(22) \quad \hat{\theta}_1 = \frac{1}{n} \sum Z_i$$

$$(23) \quad \hat{\theta}_2 = \frac{\sum Z_i W_i}{\sum W_i^2}$$

$$(24) \quad W_i = \log B_i - \frac{1}{n} \sum \log B_i$$

and $m = (n - 2)/2$. This change in m reflects the reduction in the number of degrees of freedom when two parameters are estimated. Details are given in the Appendix.

The numerical effect of using (20) can be devastating; if q_1 and q_2 truly can assume any value, then any simple time series trend in Y can be explained equally well by (i.e., you end up assigning the same posterior probability to) any population parameter values that also have a simple trend (including in the opposite direction of Y, $q_2 < 0$ case). One simple option when faced with the possibility that q_2 is less than 1.0, but almost certainly lies in some range (e.g., 0.2–1.0 for CPUE data), is to use (17) or (18) instead of (20) after replacing log B terms with q_2 log B and then integrate the resulting posterior kernel numerically over the range of q_2 values (with perhaps higher prior probabilities for some values within the range). When the Y's are relative abundances, the effect of this numerical integration will usually be to assign higher posterior probabilities to lower stock sizes (B_0).

It is not always wise to integrate over V as in (12), (18), or (20). Consider again the overall posterior function as represented in (1) for cases where more than one type of data are being used ($m > 1$, several L_k components). Retaining the variances as in (10), (17) allows tests of alternative schemes for "weighting" the data by changing V_k (high V_k means less weight on the kth type of data). Such tests are important in situations where there is fundamental uncertainty about whether or not to trust (place any weight at all upon) particular types of data such as CPUE's, especially if the $Y = qB$ assumption may have been violated by virtue of the manner in which the data were collected.

Effect of Integration on the Form of Posterior Distributions

Figure 1 shows a typical example of the differences in posterior distributions resulting from using (9)–(18). Here we generated a time series of fake relative abundance data Y_t = qB_t + normal error from a logistic population model (B_t = biomass predictions only, $B_{t+1} = B_t + rB_t(1 - B_t/B_0) - Ct$, where $r = 0.2$ and $B_0 = 1.0$). The population was subject to increasing exploitation rate over time so as to generate a catch series C_t that increased at first and then declined. Then using the same dynamic model and assuming that the correct value of r known, we generated B_t time series predictions for a range of B_0 values (from 0.1 to 2.0) along with the various sums of squares defined in the previous section. For each B_0, we calculated the posterior kernels L^*, L^q, and L^{qV}. Such calculations are easy to perform using the "Table" function in a Microsoft Excel spreadsheet[1] or analogous functions in other spreadsheets. Then the posterior probability shown for each B_0 in Fig. 1 was calculated as the corresponding L value divided by the sum of L's for all B_0 values.

Two patterns are immediately obvious, and we have found essentially the same patterns for more complex models and data series. First, for high n (many years data), the three methods converge to give the same results, as shown in Fig. 1. This is a consequence of asymptotic normality of the posterior distribution. Second, for very low n, integration over q or over q and V gives similar results: higher probabilities are assigned to lower B_0 values; in contrast, using L_k^* results in a uniform posterior for B_0 when there are few observations. Recall that use of L_k^* neglects uncertainty about the value of V. Assigning higher probabilities to lower B_0 values may seem surprising considering the assumption of a uniform prior for B_0. But when there are few data points, what those data points do provide is some constraint on B_0, of the form $\bar{Y} = qB_0$, where $\bar{Y}$ is the mean of the first few observations. Thus the posterior will be concentrated around the curve given by $\bar{Y} = qB_0$. Think of the q,B_0 plane as

[1]The authors would be pleased to provide diskettes with sample spreadsheets to anyone who needs further assistance.

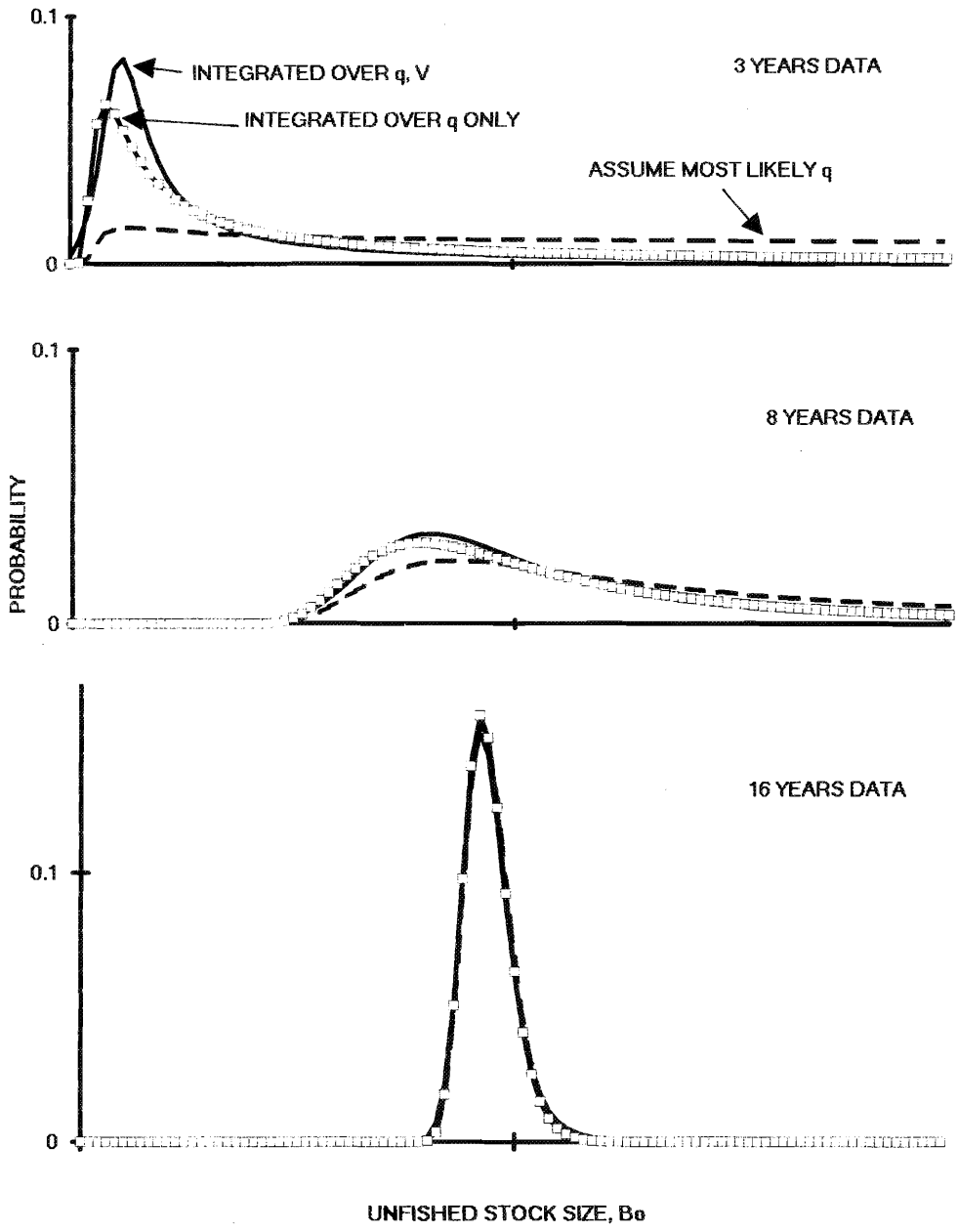

FIG. 1. Bayes posterior probability distributions for the unfished stock size parameter B_0 in a surplus production model, calculated with increasing numbers of years of data. The L^* (equation 9) calculation method ignores uncertainty in catchability q, representing each B_0 probability only for the most probable q given that B_0. The L^q and L^{qV} methods provide integral probabilities over q and the observation variance V, using the analytical expressions in equations (10) and (12).

being divided into small squares, each with an equal prior probability mass. If we integrate the posterior over q in order to obtain a marginal distribution for B_0, the total mass will be concentrated along small values of B_0, since the most of the mass of the curve lies above the small values. In other words, assigning high probabilities to low B_0 values is not a "flaw" in the analysis; rather, it is a consequence of the way the first few observations constrain the posterior distribution. Fisheries scientists who feel compelled to provide conservative assessments should actually take comfort in this result; anyone who looks at the mode of the posterior distribution (as a point estimator of B_0) will see an estimator that is biased downward, which from a conservative viewpoint is better than overestimating the stock size.

Assigning higher probabilities to low stock sizes when there are only a few relative abundance observations can be avoided either by assuming log-normality of measurement errors as in (13)–(18) or by using a prior where the logarithm of q is assumed to be uniformly distributed. In such cases the results are generally similar to the predictions of (9) in Fig. 1. The main case for using a nonconstant prior is a weakness in the argument that there are many more ways to obtain a few relative abundance values from high q, low B_0 situations than low q, high B_0 ones. In fact,

we often know a priori that q cannot be high due to the way it was measured, for example as a very small swept area chosen at random from a large area over which fish are known to be distributed.

The distributions in Fig. 1 are distinctly skewed for low n, indicating that the data allow us to confidently exclude a range of parameter "hypotheses" on one side of the most likely parameter value (distribution mode), but not on the other. In the example, we can exclude very low stock size hypotheses (could not have sustained high relative abundances given the observed catch), but not the possibility that stock size is large. This asymmetry has obvious consequences for debates about how much risk to take by allowing further increase in catches during fishery development. Such consequences emphasize the importance of not using normal approximations (based on the maximum likelihood parameter estimate and its approximate covariance matrix from nonlinear estimation) to the posterior distribution, when n is small and/or the data are uninformative.

Warnings

Bayesian methods have great potential to help improve fisheries decision-making in the face of uncertainty, if for no other reason than by bringing an end to the misunderstandings and debates triggered in management settings by point estimates and confidence limits. But they can be just as misleading as classical statistical methods, and in this section we provide warnings about three mistakes that are likely to arise in practice.

The first is a relatively simple technical warning about (9)–(20). These equations apply only to elimination of measurement parameters that appear directly in the likelihood function for the data. They will not work to integrate out the effects of nuisance parameters that appear in the dynamic model that is used to make biomass predictions, such as partial recruitment (mesh selection) factors and "process error" deviations from assumed stock–recruitment relationships. At present, calculation of marginal distributions with respect to such parameters is still a matter for massive computational effort (R. Hilborn and A. Punt, University of Washington, Seattle, Wash., personal communication) although there are promising indications of radically more efficient computational methods (J. Schnute, Pacific Biological Station, Nanaimo, B.C., personal communication; Kinas 1993).

Second, and much more important, we flatly recommend that catch/effort data never be used as a direct abundance index (assumed proportional to stock size). When catch and effort data have been gathered on a fine enough spatial scale, it should be possible to develop spatially weighted catch rate indices that are indeed reflective of population changes. But in our experience, this is not the way most CPUE series are developed, and almost all can be expected to contain misleading trends due to temporal changes in gear efficiency and increasing catchability as stock size declines ($q_2 < 1$ in (19)). Bayesian probability calculations are just as vulnerable as classical assessment estimators to the biases that such changes cause (usually overestimation of stock size and production), whether you are fitting a simple production model or tuning a massive VPA. When survey and/or good spatial catch rate data are not available, we recommend instead reliance on other indices of the intensity of exploitation, such as changes in mean body size and direct measures of total mortality rate (Csirke and Caddy 1983).

Third, we caution not to use point estimators based only on the modes of posterior distributions when setting harvest policies. Fishery development where Y remains high for a longer time (slow decrease in relative abundance) slows the convergence of the three calculations shown in Fig. 1 and causes the mode of the posterior distribution from (10)–(12) to be well below the actual B_0 for a longer period. Thus, if the modal value is used as a best estimate for setting allowable catches, then the resulting low catches will further contribute to slowing development and preventing discovery of how large B_0 actually is. This potential loss of fishery value is not a result of using Bayesian assessment procedures in general, but rather of using a point estimator rather than the full probability distribution in deciding on allowable catches and acceptable levels of risk. Thus, we caution against the use of Bayesian procedures just to provide pretty pictures of uncertainty as "high-tech" substitutes for traditional confidence limits.

Acknowledgements

We thank Ray Hilborn, Jon Schnute, and Hal Geiger for advice and direction in developing the analysis. Financial support was provided by the Natural Sciences and Engineering Research Council of Canada under grants A9239 and A5869.

References

ABRAMOWITSCH, M., AND I.A. STEGUN. 1964. Handbook of mathematical functions. National Bureau of Standards, Washington, D.C.

BERGER, J.O. 1985. Statistical decision theory and Bayesian analysis. Springer-Verlag, New York, N.Y.

BOX, G.E.P., AND G.C. TIAO. 1973. Bayesian inference in statistical analysis. Addison-Wesley, Reading, Mass.

COOKE, J.G., AND J.R. BEDDINGTON. 1985. The relationship between catch rates and abundance in fishes. IMA J. Math. Appl. Med. Biol. 2: 1–15.

CSIRKE, J., AND J.F. CADDY. 1983. Production modeling using mortality estimates. Can. J. Fish. Aquat. Sci. 40: 43–51.

DERISO, R.B. 1980. Harvesting strategies and parameter estimation for an age-structured model. Can. J. Fish. Aquat. Sci. 37: 268–282.

FOURNIER, D.A., AND A. WARBURTON. 1989. Evaluating fisheries management models by simulated adaptive control — introducing the composite model. Can. J. Fish. Aquat. Sci. 46: 1002–1012.

FRANCIS, R.I.C.C. 1992. Use of risk analysis to assess fishery management strategies: a case study using orange roughy (*Hoplostethus atlanticus*) on the Chatham Rise, New Zealand. Can. J. Fish. Aquat. Sci. 49: 922–930.

GAVARIS, S. 1988. An adaptive framework for estimation of population size. CAFSAC (Can. Atl. Fish. Sci. Adv. Comm.) Res. Doc. 88/29.

GULLAND, J.A. 1961. Fishing and the stocks of fish at Iceland. U.K. Minist. Agric. Fish. Food Invest. Ser. 2, 23(4).

GUTTMAN, I., S.S. WILKS, AND J.S. HUNTER. 1982. Introductory engineering statistics. 3rd ed. John Wiley & Sons, Inc., New York, N.Y.

HILBORN, R., E.K. PIKITCH, AND R.C. FRANCIS. 1993. Current trends in including risk and uncertainty in stock assessment and harvest decisions. Can. J. Fish. Aquat. Sci. 50: 474–480.

HILBORN, R., AND C. WALTERS. 1992. Quantitative fisheries stock assessment and management: choice, dynamics, and uncertainty. Chapman-Hall Publ. Co., New York, N.Y.

HOENIG, J.M., W.G. WARREN, AND M. STOCKER. 1994. Bayesian and related approaches to fitting surplus production models. Can. J. Fish. Aquat. Sci. 51. (In press)

JEFFREYS, H. 1961. Theory of probability. 3rd ed. Oxford University Press, London, U.K.

KINAS, P. 1993. Bayesian statistics for fishery stock assessment and management. Ph.D. dissertation, Faculty of Graduate Studies, Department of Statistics, University of British Columbia, Vancouver, B.C.

LEE, P.M. 1989. Bayesian statistics, an introduction. Edward Arnold, London, U.K.
LUDWIG, D., AND C. WALTERS. 1985. Are age-structured models appropriate for catch–effort data? Can. J. Fish. Aquat. Sci. 42: 1066–1072.
PRESS, S.J. 1989. Bayesian statistics, principles, models and applications. John Wiley & Sons, Inc., New York, N.Y.
RESTREPO, V.R., J.M. HOENIG, J.E. POWERS, J.W. BAIRD, AND S.C. TURNER. 1992. A simple simulation approach to risk and cost analysis, with applications to swordfish and cod fisheries. Fish. Bull. 90: 736–748.
SAINSBURY, K.J. 1988. The ecological basis of multispecies fisheries, and management of a demersal fishery in tropical Australia, p. 349–382. *In* J. Gulland [ed.] Fish population dynamics. 2nd ed. John Wiley & Sons, Ltd., Chichester, U.K.
SCHNUTE, J. 1985. A general theory for analysis of catch and effort data. Can. J. Fish. Aquat. Sci. 42: 414–429.
WALTERS, C.J. 1986. Adaptive management of renewable resources. Macmillan Publ. Co., New York, N.Y.

Appendix A

In this appendix, we provide details of some of the calculations, specifically those leading to (10), (12), (18), and (20). Similar derivations can be found in a variety of statistics texts, e.g., Guttman et al. (1982, p 201ff).

In order to derive (10), we begin with the formula (Abramowitsch and Stegun 1964, chap. 26)

$$\text{(A.1)} \quad \int_{-\infty}^{\infty} \exp(-u^2/2)\,du = \sqrt{2\pi}.$$

If we make the change of variables

$$\text{(A.2)} \quad u = \sqrt{I(q-q^*)/V}$$

the (A.1) becomes

$$\text{(A.3)} \quad \int_{-\infty}^{\infty} \exp(-(q-q^*)^2/(2V))\,dq = \sqrt{\frac{2\pi V}{I}}$$

which immediately implies (10).

Similarly, in order to derive (12), we begin with the definition of the Gamma function (Abramowitsch and Stegun 1964, p. 255):

$$\text{(A.4)} \quad \Gamma(p) = \int_0^{\infty} u^{p-1} \exp(-u)\,du.$$

If we set $p = -m$ and make the change of variables

$$\text{(A.5)} \quad u = -A/V$$

then (A.4) becomes

$$\int_0^{\infty} V^{-m-1} \exp(-A/V)\,dV = \frac{\Gamma(m)}{A^m}.$$

Finally, if we set $A = s^2$, the result is (12).

For the power model observation cases where

$$\text{(A.7)} \quad Y = q_1 B^{q_2}$$

we note that the likelihood function can be derived by considering an equivalent regression problem. Let

$$\text{(A.8)} \quad Z_i = \log Y_i$$

$$\text{(A.9)} \quad W_i \log B_i - \frac{1}{n}\sum \log B_i$$

$$\text{(A.10)} \quad \theta_2 = q_2$$

$$\text{(A.11)} \quad \theta_1 = \log q_1 + q_2 \frac{1}{n}\sum \log B_i.$$

In terms of these variables, we have the total sum of squares

$$\text{(A.12)} \quad \text{SS} = \sum[Z_i - (\theta_1 + \theta_2 W_i)]^2 = n(\theta_1 - \hat{\theta}_1)^2 + I(\theta_2 - \hat{\theta}_2)^2 + (n-2)s''^2$$

where

$$\text{(A.13)} \quad I = \sum W_i^2$$

$$\text{(A.14)} \quad \theta_1 = \frac{1}{n}\sum Z_i$$

$$\text{(A.15)} \quad \theta_2 = \frac{\sum Z_i W_i}{\sum W_i^2}$$

$$\text{(A.16)} \quad s''^2 = \frac{1}{n-2}\sum[Z_i - (\hat{\theta}_1 + \hat{\theta}_2 W_i]^2.$$

In this case the likelihood L is given by

$$\text{(A.17)} \quad L = V^{-n/2} \exp\left(-\frac{1}{2V}\text{SS}\right).$$

If we use flat priors on θ_1 and θ_2 (or on q_1 and q_2) and on $\log V$, then integration by the same methods as for (10) and (12), and dropping unneeded constants, gives

$$\text{(A.18)} \quad L^{q_1 q_2 V} = (s''^2)^{-m}$$

which is the result (20).