Analysis of Calibration Data for R/V Albatross IV and FSV Henry B. Bigelow

Mark S. Kaiser

Department of Statistics, Iowa State University

September 2009

1 Background

This report has been prepared as a portion of my responsibilities as a panel member for the "Vessel Calibration Analysis Review" held at the Northeast Fisheries Science Center on 11-14 August, 2009. The Chair's consensus report provides summary views of the panel, particularly as a way for NMFS to move forward on upcoming fall stock assessments. I concur with those recommendations. This individual report will expand on my own views regarding issues connected with terms of reference items d and e. These issues include methods for evaluating models relative to their representation of zero frequency tows for various species in one or both vessels, and the development of an alternative modeling strategy that may allow the estimation of calibration factors for situations (e.g., species/season combinations) resulting in fewer observations than the panel suggested were needed for such estimation in Protocol 2.2 in the Chair's consensus report. As suggested in the Chair's consensus report, the first of these might be approached through the use of simulation-based model assessment, which is described in Section 2 of this report. Of particular relevance to the assessment of zero frequencies or frequencies with which one vessel has no catch but the other some catch is the material presented in Sections 2.2.3 and 2.3.3 in what follows. A suggested framework for development of models for calibration that holds potential for dealing with the second issue, as mentioned in the Chair's consensus report in response to terms of reference item e, is then presented in Section 3. Discussion of issues needed to produce analysis using the framework developed in Section 3 is contained in Section 4, and Section 5 lays out a procedure to realize the objective of extending estimation to situations with small sample sizes. I received a request to address the issue of how length might be incorporated into this framework after preparation of this report was under way, and this is addressed in Section 6. Finally, a few summary remarks are contained in Section 7.

2 Simulation Based Model Assessment

I will cast this topic in terms of evaluating the ability of one or more models to adequately reflect the number of tows in which no individuals of a particular species were captured, but the procedures suggested could be applied to any characteristic of the observed data identified as of interest including, for example, the entire marginal distribution of counts. This material is drawn from notes that I have prepared for the course "Statistics 601: Advanced Statistical Methods" offered at Iowa State University, specifically Chapter 12.3, pages 560-581. I have attempted to renumber mathematical expressions as appropriate for this report, and remove the majority of references to specific sections and expressions that appear in previous chapters of the notes.

One of the goals of statistical modeling is to capture the key elements of a scientific mechanism in a small number of model parameters. Given the formulation of a statistical model, we may then view that model as a "data generating mechanism". If a model provides an adequate description of a set of observed data, then that (fitted) model should generate data that is similar in appearance to the actual observations. We consider here a number of methods for model assessment that rely on this notion.

2.1 Fundamental Concepts

The fundamental concept of simulation based model assessment follows directly from the view of a statistical model as a data generating mechanism. If a (fitted) model is adequate to describe an observed set of data, then data generated from that model should be similar to the observed data in all important aspects. What is needed to put this concept into practice are:

1. Data realizations from a (fitted) model.

2. A measure or measures of discrepancy between either a fitted model and a data set, or between two data sets, or a quantification of some aspect of interesting behavior in a set of data.

3. A reference distribution for the measure(s) of discrepancy chosen.

We have discussed the first item above, simulating data from a given model, previously. The second and third items are inter-twined with each other, and require additional discussion. Consideration of these aspects of simulation-based model assessment will lead to three situations; (1) a discrepancy measure between a data set and a fitted model is available but either a theoretical reference distribution is not available, or we choose not to use such a reference distribution (which is likely to be asymptotic), (2) a discrepancy measure between two data sets is available but either a theoretical reference distribution is not available or we choose not to use one if it is available, and (3) a measure of some interesting aspect of data behavior is available but we have no theoretical reference distribution available. We will consider the first two situations in the case of independent response variables. The third situation lends itself readily to either independence cases or to models with more complex dependence structures such as longitudinal settings or spatial models.

2.2 Discrepancy Measures

2.2.1 Discrepancy Between a Data Set and a Fitted Model

Most of the available quantities used as goodness of fit statistics constitute measures of discrepancy between a data set and a model fitted to the data set. A few of the more commonly used statistics for independent random variables are briefly reviewed here. The setting for all that follows in this subsection is that we have a model fitted to independent random variables $Y_1, \ldots, Y_n$ that has resulted in a set of estimated expected values $\hat{\mu}_1, \ldots, \hat{\mu}_n$, or a set of fitted probability mass or density functions $f_1(y|\hat{\boldsymbol{\theta}}), \ldots, f_n(y|\hat{\boldsymbol{\theta}})$.

1. Chi-Square Statistic.

   A traditional goodness of fit measure is the Chi-square statistic,

   $$D = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{v\hat{a}r(\hat{\mu}_i)}, \tag{1}$$

   which has, under the hypothesized model, a limiting χ^2 distribution with $n-p$ degrees of freedom where p is the number of parameters estimated.

2. Deviance.

   As was noted in the development of deviance residuals, the overall deviance given for exponential dispersion families could be used as a goodness of fit statistic. If the dispersion parameter ϕ is known (e.g., $\phi = 1$ for binomial or Poisson random components), the deviance has a limiting χ^2 distribution with $n - p$ degrees of freedom, where p is the number of estimated parameters in a model. For many models this distributional result does not hold, but the scaled deviance D^* using an estimated value of the dispersion parameter $\hat{\phi}$ still provides a measure of discrepancy between the data and a fitted model.

3. Power-Divergence Statistics.

   An entire family of goodness of fit statistics was proposed by Read and Cressie (1988) as the family of power divergence statistics. Suppose that the model under consideration is either for discrete random variables with possible values $y_i \in \{C_1, \ldots, C_k\}$, or that we have binned a model for continuous random variables into k categories $C_1, \ldots, C_k$. For $j = 1, \ldots, k$, let X_j denote the observed frequency with which the response variables $Y_1, \ldots, Y_n$ take on the value or belong to the category C_j. Suppose further that the model may be used to calculate marginal probabilities for either the possible data values (discrete case) or category membership (continuous case). In either case, let these estimated probabilities be denoted as $\hat{\pi}_j; \quad j = 1, \ldots, k$. The family of power divergence statistics is defined as, for $-\infty < \lambda < \infty$,

$$D_\lambda = \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^{k} X_j \left[\left(\frac{X_j}{n\hat{\pi}_j} \right)^\lambda - 1 \right]. \tag{2}$$

In many cases, an asymptotic χ^2 distribution is available for D_λ under the hypothesized model. This may not always be the case, however, particularly in models for continuous variables in which parameters are estimated from the density form of the model (e.g., with a likelihood or log likelihood defined in terms of densities) and D_λ is applied to categories from a subsequent binning procedure (e.g., Read and Cressie, 1998, Chapter 4.1).

The family of power divergence statistics is indexed by the parameter λ and includes a number of traditional statistics such as Pearson's Chi-square ($\lambda = 1$) and the likelihood ratio statistic for multinomial data (limit as $\lambda \rightarrow 0$). Read and Cressie (1998) suggest a generally useful value of $\lambda = 2/3$, but it seems to me that one of the strengths of the power divergence statistic is what it might reveal as λ varies. This family of statistics increases in power against "bump" alternatives as λ gets larger and positive, and increases in power for "dip" alternatives as λ gets larger and negative. A bump alternative corresponds to one or more cell frequencies substantially larger than under the hypothesized (or fitted) model, and a dip alternative corresponds to a cell frequency substantially smaller than under the hypothesized model. Thus, computing D_λ over a range of values for λ would seem to provide valuable information (see, e.g., Kaiser and Finger, 1996).

4. Kolmogorov-Smirnov Statistics.

   Consider here a set of random variables $\{Y_i : i = 1, \ldots, n\}$ that are not only independent but also identically distributed according to a theoretical distribution with density $f(y|\boldsymbol{\theta})$. Kolmogorov-Smirnov statistics are based on the empirical distribution function, defined for $-\infty < y < \infty$ as,

$$G_n(y) \equiv \frac{1}{n}\sum_{i=1}^{n} I(y_i \leq y). \tag{3}$$

Any model for *iid* random variables produces a theoretical distribution function, usually with parameters that are to be estimated. Within the context of this subsection, the estimated distribution function can be written as

$$F(y|\hat{\boldsymbol{\theta}}) = \int_{-\infty}^{y} f(t|\hat{\boldsymbol{\theta}})\, dt.$$

The Kolmogorov-Smirnov statistics are,

$$\begin{aligned} D^{+} &= \sup\left\{G_n(y) - F(y|\hat{\boldsymbol{\theta}})\right\} \\ D^{-} &= \sup\left\{F(y|\hat{\boldsymbol{\theta}}) - G_n(y)\right\} \\ D &= \max\{D^{+}, D^{-}\} \\ D' &= D^{+} + D^{-} \end{aligned} \tag{4}$$

The last quantity in (4) is often called Kuiper's statistic

A good deal of work has been conducted on determining the distributions of these and other statistics based on the empirical distribution function under various settings (see Chapter 4 of D'Agostino and Stephens, 1986, for a review). Our concern, as with the other discrepancy measures presented, will be to make use of these statistics in a simulation-based assessment procedure.

5. Cramer-von Mises Statistic.

   The Cramer-von Mises statistic is also based on the empirical distribution function (3) and is generally presented as a statistic useful in testing a hypothesized distribution F_0 as,

$$W_n^2 = n\int_{-\infty}^{\infty} \left[F_n(y) - F_0(y)\right]^2\, dF_0(y)$$

A computational form of this statistic for ordered data observations $y_{[1]} \leq y_{[2]} \leq \ldots, \leq y_{[n]}$ is

$$W_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(F_0(y_{[i]}) - \frac{2i-1}{2n} \right)^2 . \tag{5}$$

As for Kolmogorov-Smirnov and related statistics, a good deal of work has been conducted on distributional theory for the Cramer-von Mises statistic, generally in an asymptotic framework. Also as for the other statistics presented here, our concern is simply the possible use of this statistic as a measure of discrepancy.

6. Generalized Residuals.

An extremely flexible procedure for measuring the discrepancy between a set of data and a fitted model is based on the generalized residuals of expressions (12.18) in the continuous case and (12.19) in the discrete case. Under a correct and completely specified model (i.e., no estimated parameters) these residuals will behave as independent and identically distributed realizations of a uniform distribution on the interval (0, 1). Any of the goodness of fit statistics presented previously (e.g., Kolmogorov-Smirnov or Cramer-von Mises) could then be used to measure the discrepancy of these generalized residuals with a uniform distribution. With estimated parameters in a fitted model, the distribution of generalized residuals will not be uniform, and this has stymied their use in model assessment in the past. If the number of observations is large relative to the number of estimated parameters, however, the distribution of generalized residuals should be quite similar to a uniform distribution, and statistics such as those in (4) or (5) provide a useful measure of discrepancy in a simulation-based procedure.

2.2.2 Discrepancy Between Two Data Sets

Overall discrepancy between two sets of data may be quantified using two-sample versions of some of the goodness of fit statistics presented previously as discrepancy measures between a set of data and a fitted model. Among these are the two-sample versions of Kolmogorov-Smirnov statistics and the Cramer-von Mises statistic. To formalize, let $G_n(y)$ and $H_m(y)$ denote the empirical distribution functions of two sets of data, one of size n and the other of size m; in simulation-based procedures we will typically have $n = m$ but that is not strictly necessary. The two-sample Kolmogorov-Smirnov statistics are then,

$$\begin{aligned} D^+ &= \sup \{G_n(y) - H_m(y)\} \\ D^- &= \sup \{H_m(y) - G_n(y)\} \\ D &= \max\{D^+, D^-\} \\ D' &= D^+ + D^- \qquad (6) \end{aligned}$$

Let $\boldsymbol{y} = \{y_i : i = 1, \ldots, n\}$ denote the observations from one set of data, and $\boldsymbol{y}^* = \{y_j^* : j = 1, \ldots, m\}$ the observations from the other set of data. The two-sample Cramer-von Mises statistic can be written as (c.f., Conover, 1980),

$$D = \frac{mn}{(m+n)^2} \sum_{x \in \boldsymbol{y}} \sum_{x \in \boldsymbol{y}^*} \left[G_n(x) - H_m(x)\right]^2 . \qquad (7)$$

For comparison of a set of data with a fitted model the Kolmogorov-Smirnov and Cramer-von Mises statistics are typically presented in the context of independent and identically distributed random variables, as was done previously in this subsection. But any set of observed data can be used to construct a marginal empirical distribution function as given in expression (3), regardless of whether those data are assumed to have arisen from a model with *iid* random variables, independent but not identically distributed random variables, or even dependent random variables. Conditional on any factors that result in non-identical or non-independent distributions, such as covariates in a regression model, any theoretical model can be used to simulate sets of data that reflect the observed levels of those factors. This then provides a vehicle for comparison of the marginal data distribution with the marginal distribution reflected by a given fitted model.

2.2.3 Quantifying Data Behavior

In many situations we may have interest in a particular aspect of the pattern of observed data, and whether a fitted model provides a good representation of that behavior. For example, if a set of data contains a small number of extremely large values that are separated from the bulk of the observations and with fairly great spacing among themselves, we may have attempted to account for those values by using a distributional form with a long right tail in the model. We might then reasonably question whether that distributional form is sufficient to represent the observed pattern, in the size of samples we actually have. That is, a long right tail can lead to large values, but does it do so with about the correct frequency, and with the type of spacing observed in the actual data. To quantify this data pattern, we might use the difference between the largest and next-to-largest values in the data set, or we might use the average spacing among the three or four largest values.

There is great flexibility in choice of an appropriate quantification of various aspects of data patterns. The goal, of course, is to define a measure or measures that reflect behaviors we believe are important to a given problem. While general prescriptions are elusive, we can list some of the more common issues with which we might be concerned. Appropriate quantifications that reflect the aspects of data patterns listed here are largely model-specific.

1. Extreme Values.

   As illustrated immediately above, we are often concerned with data observations that fall away from modeled expected values. Even with a model we are generally satisfied with as a description of the overall data pattern we may wish to assess the frequency with which extreme observations occur. Such observations in a data set may reflect simply unusual circumstances or oddities; this is the traditional sense of data values labeled as *outliers*. But extreme observations may also reflect situations (data values) that arise somewhat infrequently, but should not be considered unusual or entirely unexpected under a given model. Our intent in assessing patterns of extreme values may be to guide model improvement, may be simply to identify an aspect of the observed situation our (fitted) model is not entirely adequate to describe, or may be to identify cases in the data that deserve closer inspection from a scientific viewpoint.

2. Unusual Data Value Frequencies.

   In some cases a set of data appear to exhibit a high relative frequency of one or two particular values. Perhaps the most common occurrence of this phenomenon is with count data having a large frequency of 0 values. We may well have modeled such a situation through use of a mixture, such as a gamma-Poisson or lognormal-Poisson mixture model. As we have seen, often the only situations in which one is able to distinguish between these two model forms are those that have a high frequency of zero values, because lognormal and gamma distributions can often be "matched" up to the first two moments except for $J-$shaped gamma distributions. A relevant question is then whether both, one, or neither of these (fitted) models has adequately captured the frequency of zero observations, or whether a more severe model, such as a two-stage model of a binary process combined with a conditional count process, is called for.

3. Need for Additional Random Terms.

It is not always clear when random effects or random data model parameters are beneficial in describing a problem. While the underlying subject matter or science can provide the strongest motivation for such terms in a model, we do not always have such guidance available. There has been, in my opinion, an unfortunate tendency among statisticians to assign the label of "overdispersion" to any situation involving large and perhaps complex patterns of variances, and to respond by including various overdispersion parameters in a model without giving interpretation to such parameters within the context of the problem. Such additional random terms in a model are not infrequently added for any number of rather arbitrarily chosen groups to "account" for overdispersion. An excellent question in many cases is whether random parameters or effects are truly needed, or whether an alternative approach to modeling variances might be preferred.

4. Mean-Variance Relations.

   Aside from those based on the normal distribution, most random model components imply a particular form of relation between expected values and variances. In some ways this can be thought of as a "systematic portion" of the random model component. It may well be possible to determine the type of mean-variance relation exhibited by a data set, and to assess potential models relative to this aspect of the observed data pattern.

5. Marginal versus Conditional Structures.

   As we have seen, many complex models involve conditioning on either data model parameters (e.g., mixed linear models, hierarchical models) or on portions of the entire set of observable random variables (e.g., Markov random field models). When this is the case we have referred to conditional and marginal model structures. In determining an appropriate quantification or quantifications of data pattern, we need to keep in mind whether the pattern or patterns we have interest in are connected with marginal or conditional model structures. A model that is fully adequate to describe a particular problem should, of course, correctly reflect both of these structures. Our ability to assess these parts of overall model structure may, however, be limited by data availability. For example, it is not uncommon in fitting Markov random field models that estimates can be obtained using maximum pseudo-likelihood or some other estimation method, and that these estimates indicate the presence of substantial dependence among the random field locations. But, if one simulates from the fitted model, it may occur that the simple average over all locations (as an estimate of the marginal mean) is no where near the observed value. One would then need to seriously question whether the data generating mechanism embodied in the fitted statistical model provides an adequate description of the actual scientific mechanisms that led to the observed data.

2.3 Simulation of Reference Distributions

Given the selection of one or more measures of discrepancy and/or quantifications of data pattern, we are prepared to simulate reference distributions for those measures or quantities. We will consider, in turn, the three situations mentioned in Section 2.1.

2.3.1 Discrepancy Between a Data Set and a Fitted Model

If the assessment is to be based on a measure of discrepancy between a the actual data and the fitted model, a value of the measure is available for the actual analysis. The chosen measure may be thought of as a functional of the estimated parameter $\hat{\theta}$ and the true but unknown parameter θ_0, so let this value of the discrepancy measure be denoted as $D(\hat{\theta}, \theta_0)$. Let the joint distribution implied by the fitted model be denoted $F(\boldsymbol{y}|\hat{\theta})$. Simulation of a reference distribution against which to assess the value $D(\hat{\theta}, \theta_0)$ may be produced by what is essentially a parametric bootstrap in the following manner:

1. For $k = 1, \ldots, M$, simulate data sets $\boldsymbol{y}^{(k)}$ from $F(\boldsymbol{y}|\hat{\theta})$.

2. For each simulated data set estimate $\hat{\theta}$ as $\theta^{(k)}$ and compute the chosen discrepancy measure as $D(\theta^{(k)}, \hat{\theta})$.

3. The empirical distribution function of the M values $\{D(\theta^{(k)}, \hat{\theta}) : \; k = 1, \ldots, M\}$ forms a reference distribution against which to assess the actual value $D(\hat{\theta}, \theta_0)$. In particular, a simulation-based $p-$value can be computed as,

$$p = \frac{1}{M} \sum_{k=1}^{M} I\{D(\hat{\theta}, \theta_0) \geq D(\theta^{(k)}, \hat{\theta})\}, \tag{8}$$

   where I is the indicator function that assumes a value of 1 if its argument is true and a value of 0 otherwise.

There is one modification of the above procedure that is worth noting if a joint sampling distribution is available for $\hat{\theta}$. The assumption inherent in this process is that the distribution of $D(\hat{\theta}, \theta_0)$ is the same as that of the $D(\theta^{(k)}, \hat{\theta})$, just as was needed in parametric bootstrap methods. This assumption can be checked, to some degree, by modifying the algorithm so that a new value of $\hat{\theta}$ is chosen from its joint sampling distribution prior to simulation of each data set $\boldsymbol{y}^{(k)}$ on which estimation of the $\theta^{(k)}$ are based. If the assumption that the distribution of a function of $\hat{\theta}$ and the true parameter θ_0 is the same as that of the function applied to estimates $\theta^{(k)}$ and the "true" $\hat{\theta}$, then this modification should not change the empirical distribution of the $D(\theta^{(k)}, \hat{\theta})$ obtained in step 3 of the procedure.

2.3.2 Discrepancy Between Two Data Sets

If the assessment is to proceed based on a measure of discrepancy between two data sets, a test quantity or test statistic is not available from only the actual data set and the model estimated from it. In this case, we need to obtain through simulation both the test quantity and its reference distribution. Let the actual data set be denoted as $\boldsymbol{y}^*$, the distribution implied by the fitted model be denoted as before by $F(\boldsymbol{y}|\hat{\theta})$, and the chosen measure of discrepancy of $\boldsymbol{y}^*$ with any other set of data be denoted as $D(\boldsymbol{y}^*, \boldsymbol{y}^{(k)})$. We assume here that $D(\boldsymbol{y}^*, \boldsymbol{y}^{(k)})$ is a summary measure that compares data sets in total, such as Kolmogorov-Smirnov or Cramer-Von Mises statistics as discussed in Section 2.2, and we assume that this measure can assume only non-negative values. A procedure to accomplish simulation-based assessment is as follows:

1. For $k = 1, \ldots, M$, simulate data sets $\boldsymbol{y}^{(k)}$ from $F(\boldsymbol{y}|\hat{\theta})$.

2. For each simulated data set, compute the discrepancy between it and the actual data as $D(\boldsymbol{y}^*, \boldsymbol{y}^{(k)})$, resulting in the set of measures $\{D(\boldsymbol{y}^*, \boldsymbol{y}^{(k)}) : k = 1, \ldots, M\}$.

3. Compute the average of these discrepancy measures as a reflection of the difference between the data and the fitted model as $T = (1/M) \sum D(\boldsymbol{y}^*, \boldsymbol{y}^{(k)})$. The statistic T will play the role of the "test statistic" for a hypothesis that the model with estimated parameter $\hat{\theta}$ provides an adequate fit to the data.

4. For each simulated data set $\boldsymbol{y}(k)$, repeat this entire process as if it were the actual data. That is, for $k = 1, \ldots, M$,

   4.1 Estimate the parameter as $\hat{\theta}^{(k)}$ from the "data" $\boldsymbol{y}^{(k)}$.

   4.2 For $j = 1, \ldots, M$, simulate data sets $\boldsymbol{y}^{(k,j)}$ from $F(\boldsymbol{y}|\hat{\theta}^{(k)})$.

4.3 For each simulated data set $\boldsymbol{y}^{(k,j)}$, compute the discrepancy between it and the "actual" data as $D(\boldsymbol{y}^{(k)}, \boldsymbol{y}^{(k,j)})$, resulting in the set of measures $\{D(\boldsymbol{y}^{(k)}, \boldsymbol{y}^{(k,j)}) : j = 1, \ldots, M\}$.

4.4 Compute the average of these discrepancy measures as $T^{(k)} = (1/M) \sum_j D(\boldsymbol{y}^{(k)}, \boldsymbol{y}^{(k,j)})$. The statistic $T^{(k)}$ plays the role of a "test statistic" for a hypothesis that the model with estimated parameter $\hat{\theta}^{(k)}$ provides an adequate fit to the "data" $\boldsymbol{y}^{(k)}$.

5. The result of step 4 is a set of values $\{T^{(k)} : k = 1, \ldots, M\}$. The empirical distribution of these M values represents a reference distribution against which to compare the actual test statistic T from step 3 of the procedure. If desired, this comparison may be represented in the form of a $p-$value as

$$p = \frac{1}{M} \sum_{k=1}^{M} I(T^{(k)} \geq T), \tag{9}$$

where $I(\cdot)$ is the usual indicator function.

There are possible modifications of this procedure that may be useful in particular situations. First, there is really no reason that M need be the same value in steps $1-3$ as it is in step 4. A modification would be to simulated $M+S$ data sets in step 1, but use only M of them in steps $2-3$. Step 4 would then be conducted for all of the $M+S$ data sets originally simulated, with M simulated data sets used throughout step 4. In steps 5 and 6, M would then be replaced with $M+S$. What is important is that the original test statistic T and the simulated test statistics $T^{(k)}$ be produced in precisely the same manner. That is, whatever is "done" to the actual data should also be "done" to each of the simulated data sets from step 1. While this modification is perfectly reasonable, it is typically not a great deal of benefit, because the most time consuming step in the procedure is step 4.1, estimating the parameter $\hat{theta}$ as $\hat{\theta}^{(k)}$ for each of the data sets simulated in step 1. It can often be difficult to automate estimation in complex models so that, for example, no individual attention to data sets is needed to determine appropriate starting values for an iterative estimation algorithm. One can contemplate any number of modifications to the procedure aimed at alleviating this difficulty by requiring no estimation beyond what is conducted with the actual data. One could, for example, compute the discrepancy measure for each of the $M(M-1)/2$ unique pairs of simulated data sets from step 1, resulting in a set of measures $\{D(\boldsymbol{y}^{(k)}, \boldsymbol{y}^{(j)}) : 1 \leq k < j \leq M\}$ from which to construct a reference distribution. Or, one might consider simulating $M+1$ data sets in step 1, and compute discrepancy measures among each of these and the remaining M sets to construct a reference distribution. Both of these possibilities would likely lead to less variability in the reference distribution than is appropriate. One could attempt to circumvent this if a sampling distribution is available for the original $\hat{\theta}$ by using it in the same manner as suggested previously in this subsection for assessing discrepancy between a data set and a fitted model. All of these, as well as other potential modifications one might consider, are entirely unevaluated at this point in time. While I cannot recommend using any of them, I do suggest that such evaluation would be a potentially profitable enterprise.

2.3.3 Quantification of Data Patterns

Suppose that model assessment is to be based on one or more given quantifications of the behavior of data that might be generated from the fitted model, as described in Section 2.3. Let $Q(\boldsymbol{y}^*)$ denote the value of such a quantity for the actual data set. Really, the only distinction between this situation and that for comparison of two data sets is that we assume $Q(\boldsymbol{y})$ can be computed from a single data set rather than requiring a pair of data sets. This does, however, simplify the procedure needed to produce a reference distribution from that required for discrepancy measures between data sets, as it eliminates the need for estimation using each data set simulated from the actual fitted model. A simulation-based procedure for arriving at a reference distribution for $Q(\boldsymbol{y}^*)$ can be outlined as follows.

1. For $k = 1, \ldots, M$, simulate data sets $\boldsymbol{y}^{(k)}$ from $F(\boldsymbol{y}|\hat{\theta})$.

2. For each simulated data set, compute the quantity $Q(\boldsymbol{y}^{(k)})$, resulting in the set of quantities $\{Q(\boldsymbol{y}^{(k)} : k = 1, \ldots, M\}$.

3. A simulation-based $p-$value for $Q(\boldsymbol{y}^*)$ may then be computed as

$$p = \sum_{k=1}^{M} I\left(|Q(\boldsymbol{y}^*)| \geq |Q(\boldsymbol{y}^{(k)})|\right), \tag{10}$$

   where I is the usual indicator function.

If the quantity $Q(\boldsymbol{y})$ can assume only non-negative (or non-positive) values, such as the proportion of zero values in a model for count data, the absolute values of expression (10) are not needed. As for the simulation of reference distributions for measures of discrepancy between a data set and a fitted model, one might choose to simulate data sets in step 1 after having first chosen a parameter value from the sampling distribution of $\hat{\theta}$ if one is available.

2.4 Issues in Model Assessment

In keeping with the theme of this entire set of notes, this chapter has attempted to present model assessment within the context of general approaches rather than as a recipe book for particular models. In this context there are a number of larger issues involved in model assessment that deserve mention. The resolution of these issues in any particular situation must involve the objectives of a statistical analysis, as discussed in the introductory portion of this chapter. Here, we identify and briefly discuss several of the major issues.

2.4.1 Goodness of Fit as a Statistical Test

One major issue that arises in consideration of model assessment is whether it is possible to test the "goodness of fit" of a model within the framework of traditional statistical hypothesis testing. Actually, this is something of a misstatement as this really is not an "issue"; nearly all statisticians agree that the answer is no it is not possible to cast goodness of fit in the framework of Neyman-Pearson hypothesis tests. But the reasons for this do raise issues as to how the results of goodness of fit tests should be interpreted.

We are familiar with tests of the form $H_0 : \mu = \mu_0$ versus the disjunctive alternative $H_1 : \mu \neq \mu_0$, which may also be taken to mean "something other than μ_0". Similarly, traditional goodness of fit tests are often formulated as $H_0 : \boldsymbol{Y} \sim F(\boldsymbol{y}|\theta_0)$ versus $H_1 : \boldsymbol{Y}$ has some other distribution. There are a number of differences, however, between testing the goodness of fit of a model and the usual development of hypothesis tests about values of parameters within an assumed model.

1. Hypothesis tests are generally developed by first considering simple null and simple alternative hypotheses for the value of a parameter, such as $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, where θ_0 and θ_1 are specified numerical values. In this setting, the Neyman-Pearson lemma can be used to find most powerful tests of specified size. Often, it can be shown that most powerful tests do not depend on the specific value θ_1, making them most powerful for composite alternative hypotheses, at least for one-sided alternatives (e.g., $H_1 : \theta > \theta_0$). This development simply cannot be pushed through in the case of goodness of fit hypotheses. Even when the null hypothesis is simple, the alternative will be not only composite, but also vague. That is, the composite alternative hypothesis in a goodness of fit setting cannot be generated as a known class of simple alternatives in the same way as composite alternatives for parameters can be in the classic hypothesis testing theory as, for example, $H_1 : \mu > \mu_0$,

and there is no reason to even consider taking the step to unbiased tests that allow two-sided alternatives. So, no concept of optimality is available for goodness of fit tests.

2. The null hypothesis in a goodness of fit test is typically composite itself, such as $H_0 : \boldsymbol{Y} \sim N(\mu, \sigma^2)$ with unspecified parameters μ and σ^2. This removes goodness of fit tests one additional step from the theory used to develop hypothesis tests that have known (and in some situations optimal) properties.

3. While the previous comments are true, they may also be largely moot, because the typical interest in a goodness of fit test is to "accept" the null hypothesis. This is directly opposed to the logical development of hypothesis testing (c.f., the nested syllogism of experimentation in Chapter 4.2) and the concept that failing to reject the null hypothesis is not the same as accepting the null hypothesis.

The conclusion that must be reached is that goodness of fit tests cannot be considered tests of hypotheses in the usual sense. This should not be seen as any great loss. We understand that a statistical model is a conceptualization of the situation under investigation so that testing whether a model is "true" is not a viable enterprise in the first place. One view is that we can still interpret a $p-$value that results from a test-like procedure (e.g., expression 12.29, 12.30, or 12.31) as a measure of "surprise" under the assumption that the fitted model is an adequate description of the mechanisms that produced the observed data. This may be possible, but even this needs to be qualified in some cases, in part because of the next issue to be discussed.

2.4.2 Goodness of Fit Versus Model Selection

A fundamental issue that has resulted in some lengthy discussions in the literature is whether any model assessment can truly take place in the absence of an potential alternative model, regardless of whether that alternative is available in explicit or only implicit form. Two aspects of this issue seem rather immediate. First, if we have an alternative model explicitly available, we most likely will conduct some type of a test or comparison of the two models to select between them. Second, we would certainly desire to be able to assess a model in the absence of any posited alternative. After all, the model under consideration at this point may well have already been selected from a set of possible alternatives, assuming the status of what we sometimes refer to as a "final model" in an analysis. We would like to be able to declare our final model as adequate, or indicate that even our best efforts have not produced a satisfying result and send the problem back to the "scientific drawing board". I have called this issue fundamental because the question is whether we can actually achieve this objective within the context of what is known about statistical analysis, not whether it would be a good idea were it possible.

At the heart of the position that it is not truly possible to assess a model without some type of alternative in mind is the fact that all, or nearly all, statistical test statistics are more sensitive to some types of discrepancy between data and model (or two data sets) than they are to others. For example, in Section 2.2.1 we described a parameterized family of test statistics called power divergence statistics. This family includes traditional Pearson Chi-square and likelihood ratio statistics. Depending on the value of the parameter λ in expression (2), this statistic is more sensitive to particular alternative data patterns such as "bump" (one or two large frequencies) or "dip" (one or two small frequencies) alternatives. As a result, the value of λ chosen (even if that value results in Chi-square or likelihood ratio statistics) implies at least an implicit alternative. Given this, the argument can be made that our efforts would be better directed toward explicitly formulating an alternative model and using some form of model comparison between it and the model under consideration. This argument certainly seems to carry some force in the situation in which a fitted model is being compared to a data set using statistics such as those described in Section 2.2.1. At the very least, the indication is that we should not pretend that such statistics quantify the “adequacy” of a fitted model relative to a set containing equally weighted members of “anything else”.

If our chosen measure of discrepancy is based on a particular feature of the data or quantification of a data pattern, one view is that the selection of that feature really has defined the type of alternative being considered. For example, if our measure is the proportion of 0 counts in a data set, then the unspoken alternative to our fitted model is one that produces data with the relative frequency of zeros as observed. To call this an alternative could also, however, be considered a tautology (e.g., the alternative is nothing more than the data as so is redundant and without additional meaning). This is my view, and I would claim that the question being asked in such an assessment (could this model have reasonably generated these data) comes as close to being “alternative free” as we can achieve.

3 A Class of Hierarchical Models for Calibration

In this section of the report I present an initial formulation of a class of hierarchical models that might allow estimation for particular situations (e.g., species or species/season combinations) that lack a large amount of data. It is possible that this strategy could also provide a unified structure within which to approach estimation of calibration factors in general.

3.1 General Principles

I first reiterate a point made at the review that the selection of situations for which calibration factors are to be estimated, and the inclusion or exclusion of model terms (e.g., covariates) should be based on a combination of two considerations, biological sensibility and observed data pattern. The need to rely on both of these considerations, rather than primarily one or the other, is accentuated by the limited time in which the Albatross and the Bigelow could be operated in tandem. Given the number of possible effects that could be considered, reliance on only indications of statistical "significance" in this setting would almost certainly lead to models that incorporate spurious factors. On the other hand, the fundamental principle that pronouncements of various factors as "important" should be verified through empirical observation indicates that selection of models based only on what we believe we "know" is not a tenable scientific option. Given the unfortunate combination of budgetary realities and mechanical difficulties that restricted the available data for comparison of the Albatross and Bigelow, the overall goal should be identification of models that provide stability in prediction of what one vessel would catch based on observation of the other. That such stability in prediction does not always correspond to the model having the "best fit" for a particular data set (as indicated for example by AIC values) is illustrated in an entirely different context by Kaiser and Finger (1996). The stability being sought in the current problem must come primarily from widely accepted biological understanding.

The strategy proposed here to develop a class of generally applicable calibration models is that of hierarchical modeling, in which calibration factors for individual situations (e.g., a species/season or species/length class or species/season/sex set of data) are estimated by "borrowing strength" from observation of other similar situations. Although such models are not inherently Bayesian, a Bayesian approach to estimation and inference is a natural statistical choice and will be advocated here.

Our objective is to estimate the ratio in “catchability” of the two vessels for a set of situations. Differences caused by specifics of gear, acoustic factors, towing speeds, and other relevant factors are subsumed under the heading of vessel differences. Considered as a function of vessel operation and fish behavior (for a particular situation) catchability should conceptually be a fixed value, and will be represented as such in the models proposed. Variability in the density of fish encountered by the two vessels is accounted for through the use of “mixing distributions” or “random parameter models” in the overall statistical structure. The class of models proposed contains three main components, the “data model”, the “mixing distribution(s)”, and the “prior distribution(s)”. These are considered in turn.

3.2 The Data Model

Following the majority of models proposed for estimation of calibration factors, we will assume that the number of fish caught in an individual tow can be represented as corresponding to an observation of a random variable having a Poisson probability mass function. To this end, let $Y_{A,i}$ be a random variable corresponding to the number of fish caught by the Albatross at sampling station i and $Y_{B,i}$ similarly correspond to the number of fish caught by the Bigelow at station i. Let $D_{A,i}$ denote the density of fish encountered by the Albatross and $D_{B,i}$ denote the density of fish encountered by the Bigelow, both at station i. Let $V_{A,i}$ denote the volume (or, if appropriate for a situation, the area) of water sampled by the Albatross and $V_{B,i}$ the volume sampled by the Bigelow, also both at station i. Here, $Y_{A,i}$ and $Y_{B,i}$ are random variables associated with observable phenomena, $D_{A,i}$ and $D_{B,i}$ are random variables associated with unobservable phenomena, and $V_{A,i}$ and $V_{B,i}$ are constants that can be calculated based on tow duration and speed. Assume that, given $D_{A,i}$ and $D_{B,i}$, $Y_{A,i}$ and $Y_{B,i}$ have independent Poisson distributions with parameters $q_A V_{A,i} D_{A,i}$ and $q_B V_{B,i} D_{B,i}$, respectively. That is, assume that the probability mass functions of $Y_{A,i}$ and $Y_{B,i}$ are,

$$\begin{aligned} f(y_{A,i}|D_{A,i}, q_A) &= \frac{1}{(y_{A,i})!} \{q_A V_{A,i} D_{A,i}\}^{y_{A,i}} \exp\{-q_A V_{A,i} D_{A,i}\} \\ f(y_{B,i}|D_{B,i}, q_B) &= \frac{1}{(y_{B,i})!} \{q_B V_{B,i} D_{B,i}\}^{y_{B,i}} \exp\{-q_B V_{B,i} D_{B,i}\} \qquad (11) \end{aligned}$$

In expression (11) q_A and q_B are the catchabilities of the Albatross and the Bigelow for the specified situation, respectively. Different versions of the model are produced through different assumptions about the densities of fish encountered by the two vessels, $D_{A,i}$ and $D_{B,i}$.

3.3 Mixing Distributions

Work conducted by NMFS and presented at the review indicates that models incorporating a stochastic component over tows or sampling stations have superior performance to models that do not incorporate such a component. This is not surprising given the restrictive nature of mean-variance relation in Poisson distributions. Three basic models (meaning without covariates such as length) are given by three different assumptions about the densities of fish a a station, although all of these models take densities to be random quantities. Assume that, for a set of stations to be considered in estimation of a calibration factor, $i = 1, \ldots, S$ say, $Y_{A,i}$ and $Y_{B,i}$ have probability mass functions given in expression (11).

3.3.1 Model 1.

Assume that, for $i = 1, \ldots, S$, $D_{A,i} = D_{B,i} = D_i$. Let $D_i \sim iid$ according to some distribution that has density $g(d_i|\boldsymbol{\theta})$ with support on the positive line and parameter $\boldsymbol{\theta}$. Assume that $Y_{A,i}$ and $Y_{B,i}$ are conditionally independent given $D_i = d_i$. The joint conditional probability mass function of the two catches is then

$$f(y_{A,i}, y_{B,i}|d_i, q_A, q_B) = f(y_{A,i}|d_i, q_A)\, f(y_{B,i}|d_i, q_B) \qquad (12)$$

and the marginal model becomes

$$h(y_{A,i}, y_{B,i}|q_A, q_B, \boldsymbol{\theta}) = \int_0^\infty f(y_{A,i}, y_{B,i}|d_i, q_A, q_B)\, g(d_i|\boldsymbol{\theta})\, d(d_i), \tag{13}$$

where the notation $d(d_i)$ is just for clarity because of the multiple "d"s; expression (13) is an ordinary Riemann integral. Depending on the choice of $g(d_i|\boldsymbol{\theta})$, it may or may not be possible to evaluate the integral (13) analytically. If not, this difficulty can be overcome through inclusion of the densities $d_i : i = 1, \ldots, S$ as "parameters" in an overall MCMC algorithm. If (13) can be relatively easily evaluated we may or may not choose to do so, as will be discussed in Section 4. Also note that, while $Y_{A,i}$ and $Y_{b,i}$ are assumed conditionally independent, they will not be marginally independent. Assuming independence among sampling stations, the likelihood for this model is

$$h(\boldsymbol{y}_{A,i}, \boldsymbol{y}_{B,i}|q_A, q_B, \boldsymbol{\theta}) = \prod_{i=1}^{S} h(y_{A,i}, y_{B,i}|q_A, q_B, \boldsymbol{\theta}). \tag{14}$$

3.3.2 Model 2

In this model we assume that $D_{A,i} \neq D_{B,i}$, but that values of these random quantities can be considered as arising as two independent draws from the same mixing distribution. So $D_{A,i}$ has density $g(d_{A,i}\boldsymbol{\theta})$ and $D_{B,i}$ has the same density $g(d_{B,i}|\boldsymbol{\theta})$. In this case, $Y_{A,i}$ and $Y_{B,i}$ are both conditionally and marginally independent with marginal probability mass functions

$$\begin{aligned} h(y_{A,i}|q_A, \boldsymbol{\theta}) &= \int_0^\infty f(y_{A,i}|q_A, d_{A,i})\, g(d_{A,i}|\boldsymbol{\theta})\, d(d_{A,i}) \\ h(y_{B,i}|q_B, \boldsymbol{\theta}) &= \int_0^\infty f(y_{B,i}|q_B, d_{B,i})\, g(d_{B,i}|\boldsymbol{\theta})\, d(d_{B,i}). \end{aligned} \tag{15}$$

The joint marginal data model for $Y_{A,i}$ and YB, i at sampling station i is given as the product $h(y_{A,i}, y_{B,i}|q_A, q_B, \boldsymbol{\theta}) = h(y_{A,i}|q_A, \boldsymbol{\theta})\, h(y_{B,i}|q_B, \boldsymbol{\theta})$, and the likelihood becomes

$$h(\boldsymbol{y}_A, \boldsymbol{y}_B|q_A, q_B, \boldsymbol{\theta}) = \prod_{i=1}^{S} h(y_{A,i}, y_{B,i}|q_A, q_B, \boldsymbol{\theta})$$

$$= \prod_{i=1}^{S} h(y_{A,i}|q_A, \boldsymbol{\theta})\, h(y_{B,i}|q_B, \boldsymbol{\theta}). \tag{16}$$

3.3.3 Model 3

Here, assume that $D_{A,i} \neq D_{B,i}$ as in Model 2, but that the observed values of these variables can be represented as correlated draws from a bivariate mixing distribution with density $g(d_{A,i}, d_{B,i}|\boldsymbol{\theta})$, with support on the strictly positive quadrant of $\mathcal{R}^2$ and parameter $\boldsymbol{\theta}$. The joint conditional probability mass function for the two catches now becomes

$$f(y_{A,i}, y_{B,i}|d_{A,i}, d_{B,i}, q_A, q_B) = f(y_{A,i}|d_{A,i}, q_A)\, f(y_{B,i}|d_{B,i}, q_B) \tag{17}$$

and the marginal model has the form
$h(y_{A,i}, y_{B,i}|q_A, q_B, \boldsymbol{\theta}) =$

$$\int_0^\infty \int_0^\infty f(y_{A,i}, y_{B,i}|d_{A,i}, d_{B,i}, q_A, q_B)\, g(d_{A,i}, d_{B,i}|\boldsymbol{\theta})\, d(d_{A,i})\ d(d_{B,i}) \tag{18}$$

With the distinction between marginal probability mass functions in expressions (13) and (18) noted, the likelihood for this model again has the form of expression (14).

3.3.4 Comparison of Models

A variety of specific models can be developed through selection of particular forms for the mixing distributions $g(d_i|\boldsymbol{\theta})$ in Model 1, the common $g(d_{A,i}|mbtheta)$ and $g(d_{B,i}|\boldsymbol{\theta})$ in Model 2, and $g(d_{A,i}, d_{B,i}|\boldsymbol{\theta})$ in Model 3. For Models 1 and 2, choices that readily suggest themselves are gamma, generalized gamma, lognormal, and extreme value distributions. Any of these choices could be folded into a finite mixture to produce greater frequencies of zero catch if that appears necessary. For Model 3 the choices are more limited but include bivariate lognormal and (possibly) extreme value distributions. One could, of course include truncated (at zero) normal distributions for each model, but it seems reasonable that something with longer right tail behavior would be more appropriate. Each of the three model structures conceptualizes the number of fish captured by the Albatross and Bigelow as arising from somewhat different probabilistic mechanisms, and there are also some analogies with models considered by NMFS and presented at the review.

1. Model 1.

   Model 1 takes the density of fish encountered by the two vessels at a sampling station to be the same. While the catch of the Albatross and Bigelow are taken to be conditionally independent in this model, they will not be marginally independent because of this common value of d_i. This might be most appropriate for situations involving species that are widespread and are either non-schooling or form schools that are structured horizontally. A question with the structure of Model 1 is whether it is flexible enough to produce the level of correlations observed between catches of the two vessels. There is similarity between models of this structure and what was presented at the review as a "correlated negative binomial" model. These structures may even be equivalent if the mixing distribution for D_i is taken to be gamma, although I have not worked out the details. As already pointed out, however, the more general structure of Model 1 allows any number of distributions to be considered for the D_i within the same framework.

2. Model 2.

   Model 2 allows the densities of fish encountered by the two vessels at a sampling station to differ, but takes these quantities to be independent. As a result, the catches of the vessels will be independent in the marginal data model as well. Evidence was presented at the review that any number of species show marked correlation in catch across sampling stations, and Model 2 would most likely not be fully adequate to describe such situations. But there may be cases in which correlation among vessels is weak, and these are the situations for which one might consider the use of this model. Situations involving species that have extremely patchy distributions or schools that are vertically structured might lend themselves to representation by Model 2. There is similarity of Model 2 with the "independent negative binomial" model presented at the review but, as for Model 1, this would be most true only if the mixing distributions were taken as gamma. The beta-binomial model presented at the review is also similar to Model 2 if the mixing distribution is gamma and the additional step is taken of conditioning on the sum $D_{A,i} + D_{B,i}$.

3. Model 3.

   Like Model 2, Model 3 allows the densities of fish encountered by the Albatross and Bigelow to differ but, unlike Model 2, attempts to maintain some degree of correlation across sampling stations. This model is an attempt to incorporate the "best" features of both Model 1 and Model 2 into one structure. While somewhat more elegant than either of the previous models, a drawback of this model is difficulty in constructing an appropriate bivariate mixing distribution for $(D_{A,i}, D_{B,i})$. This difficulty would be accentuated if a need arises to make use of a finite mixture formulation of to account for large numbers of 0 catches. That is, the model reflects 0 catches through the combination of low density values and the frequency for 0 values resulting from a Poisson observation model with small mean (e.g., $q_A V_{A,i} D_i$). If this structure is not capable of reflecting the observed frequencies of zero catch (as assessed using methods similar to those outlined in Section 2 of this report, for example) then the natural option is to model fish densities as arising from a finite mixture of a point mass at 0 and a distribution for non-zero fish density. While doing so would seem possible for both Model 1 and Model 2, achieving this in a bivariate setting for the pair $D_{A,i}, D_{B,i}$ would prove to be a challenge. This presents an interesting research question, but is perhaps a bit beyond what could be expected in an analysis that is practical at the current time.

3.3.5 Considerations in Model Selection

Three aspects that should be considered in choosing any of model structures, or any additional alternatives, are as follows.

1. The frequency of 0 values in marginal distributions of $Y_{A,i}$ and $Y_{B,i}$.

2. The behavior in the right tail of marginal distributions of $Y_{A,i}$ and $Y_{B,i}$.

3. Correlation between values of $Y_{A,i}$ and $Y_{B,i}$ in the marginal distribution.

Although I was not necessarily of this opinion at the review itself, further reflection causes me to suggest that the first two of these might be more important to adequately model than the third, although correlation should not be ignored entirely. This is because we would expect the correct reflections of frequencies of various levels of catch to have a greater effect on point estimation of q_A and q_B, while correlation would be expected to effect primarily the precision of estimation and a proper quantification of uncertainty. Thus, I would recommend focusing first on models having the structures of Model 1 or Model 2.

3.3.6 The Beta-Binomial Formulation

The beta-binomial model advocated by NMFS at the review introduces an interesting question relative to the formulation and selection of models to be considered. That model seemed to have generally acceptable behavior in the simulations conducted, and the question is what might produce this phenomenon relative to the other models considered. Consideration of the particular mechanisms used to generate data for the simulation studies is certainly appropriate, but it strikes me that there may be a more fundamental issue involved. Casting the observation process in terms of a binomial results from conditioning on the total catch at each sampling station. Such conditioning intuitively removes a major source of variability from the problem, that being the overall density of fish being sampled. Any differences in $V_{A,i}$ and $V_{B,i}$ as well as $D_{A,i}$ and $D_{B,i}$ are subsumed by now non-constant q_A and q_B (which could then be taken as $q_{A,i}$ and $q_{B,i}$), and variability in the ratio of these quantities is conditioned on the sum $Y_{A,i} + Y_{B,i}$. This strikes me as a less than fully pleasing conceptualization of the problem, but it may well present a useful vehicle for estimation of q_B/q_A in terms of an expected binomial parameter. Thus, I would not recommend that NMFS completely abandon consideration of this formulation. One point to be made is that the strategy for borrowing information across similar situations (e.g., species groups) as described in Section 5 of this report could be applied to the beta-binomial model as well as to the models presented previously, and details of assigning prior distributions to this model will be considered in the next subsection along with the other models. At the same time, a single procedure that produces calibration factors in both directions (Albatross to Bigelow and Bigelow to Albatross) will not result from using the beta-binomial model.

3.3.7 Example: Gamma Mixing Distribution for D_i in Model 1

As an example in which the integrals of expression (13) can be evaluated analytically, consider Model 1 with $g(d_i|\boldsymbol{\theta})$ taken to be the density of a gamma distribution. Let $\boldsymbol{\theta} \equiv (\alpha, \beta)$, and assume that

$$g(d_i|\alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\, d_i^{\alpha-1}\, \exp(-\beta d_i); \;\; d_i > 0. \qquad (19)$$

The joint data model of expression (12) is,

$$f(y_{A,i}, y_{B,i}|q_A, q_B, d_i) =$$

$$\frac{1}{y_{A,i}!y_{B,i}!}(q_A V_{A,i})^{y_{A,i}}(q_B V_{B,i})^{y_{B,i}} d_i^{y_{A,i}+y_{B,i}} \exp\{-(q_A V_{A,i} + q_B V_{B,i})d_i\} \qquad (20)$$

Then using (19) and (20) in the integral of expression (13), the marginal joint probability mass function for $Y_{A,i}$ and $Y_{B,i}$ becomes

$$\begin{aligned} h(y_{A,i}, y_{B,i}|q_A, q_B, \boldsymbol{\theta}) &= \frac{\beta^\alpha (q_A V_{A,i})^{y_{A,i}}(q_B V_{B,i})^{y_{B,i}}}{\Gamma(\alpha) y_{A,i}! y_{B,i}!} \\ &\quad \times \int_0^\infty d_i^{\alpha+y_{A,i}+y_{B,i}-1} \exp\{-(q_A V_{A,i} + q_B V_{B,i} + \beta)d_i\}\ d(d_i) \\ &= \frac{\beta^\alpha (q_A V_{A,i})^{y_{A,i}}(q_B V_{B,i})^{y_{B,i}} \Gamma(\alpha + y_{A,i} + y_{B,i})}{\Gamma(\alpha) y_{A,i}! y_{B,i}! (q_A V_{A,i} + q_B V_{B,i} + \beta)^{\alpha+y_{A,i}+y_{B,i}}} \\ &= \frac{\beta^\alpha (q_A V_{A,i})^{y_{A,i}}(q_B V_{B,i})^{y_{B,i}} \left[\prod_{h=0}^{y_{A,i}+y_{B,i}-1} (\alpha + h)\right]}{y_{A,i}! y_{B,i}! (q_A V_{A,i} + q_B V_{B,i} + \beta)^{\alpha+y_{A,i}+y_{B,i}}} \qquad (21) \end{aligned}$$

Although expression (21) appears somewhat formidable at first glance, forming a joint over sampling stations as in (14) and taking the logarithm would result in a quite manageable set of summations that would not be difficult to compute, along with the necessary derivatives, if maximum likelihood estimation was desired. A likelihood analysis might be interesting and could provide preliminary indications of model behavior. But likelihood analysis does not lend itself to the objective of borrowing strength across similar species for the purpose of producing calibration factors for individual species with low to moderate sample sizes. For this we will need a Bayesian approach and prior distributions to be discussed presently.

3.4 Prior Distributions

In the general structure of any of the models presented previously (aside from the beta-binomial which will be considered at the end of this subsection), prior distributions are needed for q_A, q_B, and $\boldsymbol{\theta}$, although the composition of $\boldsymbol{\theta}$ will differ between models. I will focus here on the structure of Model 1. A natural overall prior structure is the product form

$$\pi(q_A, q_B, \boldsymbol{\theta}) = \pi_A(q_A)\,\pi_B(q_B)\,\pi_\theta(\boldsymbol{\theta}). \tag{22}$$

Prior distributions for q_A and q_B should have support on the interval (0, 1) and the obvious choice is to use beta distributions for this purpose. The parameters of these beta distributions could well be selected to result in uniform priors for these two quantities of primary interest. That is, $\pi_A(\cdot)$ and $\pi_B(\cdot)$ in (22) could be taken as

$$\begin{aligned} \pi_A(q_A) &= 1; \quad 0 < q_A < 1 \\ \pi_B(q_B) &= 1; \quad 0 < q_B < 1 \end{aligned} \tag{23}$$

The dimension of $\boldsymbol{\theta}$ may vary depending on the mixing distribution chosen for $g(d_i|\boldsymbol{\theta})$, and appropriate prior distributions for these components will vary as well. The more obvious choices for g have two parameters, such as α and β in the gamma formulation or μ and σ^2 in a lognormal formulation or what are sometimes denoted as ψ and ϕ in an extreme value model (ψ is a location parameter and ϕ a scale parameter in this distribution). In general, a product form can be applied to the components of $\boldsymbol{\theta}$ just as to the full parameter vector in expression (22). Some natural choices for priors with two example models follow.

1. Gamma Formulation.

   Here, it would be natural to take $\pi(\alpha, \beta) = \pi_\alpha(\alpha)\,\pi_\beta(\beta)$ with $\pi_\beta(\beta)$ being diffuse gamma (to exploit conditional conjugacy) and $\pi_\alpha(\alpha)$ either improper, or proper on a large interval of the positive line.

2. Lognormal Formulation.
   For a model in which $g(d_i|\boldsymbol{\theta})$ is taken to be lognormal with $\boldsymbol{\theta} \equiv (\mu, \sigma^2)$ we might take $\pi(\mu, \sigma^2) = \pi_\mu(\mu)\,\pi_\sigma(\sigma^2)$ in which $\pi_\mu(\mu)$ is normal with a large variance (e.g., 1000 or 5000) and $\pi_\sigma(\sigma^2)$ is either proper uniform on a large interval or diffuse inverse gamma (say with parameters 0.01 and 0.01).

Choosing priors for models with more extreme tail behavior in the mixing distribution for fish densities, such as extreme value or generalized gamma distributions might be a bit more involved only because we have less experience with such models. I have suggested some priors for a model with extraordinary right tail behavior in work on another project involving discard estimation, and something along those lines might apply to this problem as well, if it is determined that more standard formulations for $g(d_i|\boldsymbol{\theta})$ are not fully adequate.

Assigning priors to a beta-binomial model is actually fairly straightforward. From a standard (α, β) parameterization of a beta distribution, let $\mu = \alpha/(\alpha+\beta)$ be the expected value and let $\phi = 1/(\alpha+\beta+1)$ be an additional dispersion parameter. In this parameterization, the sets of possible values for μ and ϕ are both the unit interval (0, 1). This suggests uniform or some other beta distribution for individual priors $\pi_\alpha(\alpha)$ and $\pi_\beta(\beta)$, with the joint prior in product form, $\pi(\alpha, \beta) = \pi_\alpha(\alpha)\,\pi_\beta(\beta)$.

4 Simulation of Posteriors

It can be taken as a given that posterior distributions will not be available in closed form, and that Markov Chain simulation will be needed to produce samples from the joint (and, hence also marginal) posteriors of q_A, q_B and $\boldsymbol{\theta}$. The most likely overall structure for a Markov Chain Monte Carlo algorithm in this problem would seem to be that of the Gibbs Sampler. I assume throughout that the reader has a basic familiarity with the Gibbs Sampling algorithm. If this is not the case, then much of what is presented in this subsection will be without meaning. For a Gibbs algorithm what is needed are the full conditional posteriors or functions that are proportional to them. There are two situations that might arise, that in which the marginal data model (e.g., expression 13) is available in closed form and that in which it is not. I consider these two situations in turn.

4.1 Closed Form Data Model

If $h(y_{A,i}y_{B,i}|q_A, q_B, \boldsymbol{\theta})$ can be derived in closed form, such as in expression (21) for the gamma model example, then there are only a relatively small number of conditional posteriors that will be needed. Suppose that $\boldsymbol{\theta} = (\theta_1,\, \theta_2)^T$. Let $\boldsymbol{y}$ denote the entire set of observed catches from both the Albatross and the Bigelow. The distributions that will be required for simulation are then $p(q_A|q_B, \theta_1, \theta_2, \boldsymbol{y})$, $p(q_B|q_A, \theta_1, \theta_2, \boldsymbol{y})$, $p(\theta_1|q_A, q_B, \theta_2, \boldsymbol{y})$ and $p(\theta_2|q_A, q_B, \theta_1, \boldsymbol{y})$. All of these conditional posteriors are proportional to the joint posterior which, in turn can be represented by

$$\pi(q_A, q_B, \theta_1, \theta_2, \boldsymbol{y}) \propto h(\boldsymbol{y}_A, \boldsymbol{y}_B|q_A, q_B, \theta_1, \theta_2)\, \pi_A(q_A)\, \pi_B(q_B)\, \pi_1(\theta_1)\, \pi_2(\theta_2). \tag{24}$$

Conditional posteriors (un-normalized) result from considering the right hand side of (24) as functions of the individual parameters for fixed values of all other quantities involved. Typically, when using a marginal data model there is not a great deal of simplification available for various conditional posteriors because $h(\boldsymbol{y}_A, \boldsymbol{y}_B|q_A, q_B, \boldsymbol{\theta})$ is a product given by expression (14). It is not uncommon that this can result in computational difficulties such as overflow (e.g., an expression exceeding exp(360)). Sometimes such problems can be overcome through computational maneuvers (e.g., a product is the exponentiation of the sum of its log terms) but other times computations can become prohibitive, and one might choose to consider the problem in the context of a situation in which a closed form is not available for the data model. The potential benefit of using a marginal closed form data model, if such is available, is that the number of conditional posteriors that must be sampled from in an overall Gibbs algorithm is greatly reduced. The potential drawback is that computation may be difficult or impossible.

One possible strategy for sampling from the conditional posteriors is the use of a Metropolis rule. If this is the case, it may be beneficial to sample q_A and q_B jointly, using the acceptance rule on this bivariate pair rather than sampling of the two quantities individually. The overall algorithm structure would still be that of a Gibbs Sampler, but the steps of sampling from $p(q_A|q_B, \theta_1, \theta_2, \boldsymbol{y})$ would be folded into one step of sampling from the bivariate distribution $p(q_A, q_B|\theta_1, \theta_2, \boldsymbol{y})$. A natural proposal distribution to use in this scenario would be independent beta (or uniform) distributions for q_A and q_B, resulting in a random walk Metropolis embedded in an overall Gibbs algorithm.

4.2 Data Model with Unevaluated Integrals

If $h(y_{A,i}, y_{B,i}|q_A, q_B, \boldsymbol{\theta})$ cannot be determined analytically because the integral in (13) is intractable, the usual approach to simulation from the conditional posteriors is to effectively conduct the integration as a portion of the simulation procedure. For Model 1, this would be accomplished by defining the fish densities D_i; $i = 1, \ldots, S$ to be additional "parameters" in the model. Continue to assume that $\boldsymbol{\theta} \equiv (\theta_1, \theta_2)^T$. Let $\mathcal{D} \equiv \{d_i : i = 1, \ldots, S\}$ be defined as the set of fish densities for all sampling sites, and let $\mathcal{D}_{-i} \equiv \mathcal{D} \backslash d_i$ be defined as this set less the value at sampling station i. Then the conditional posteriors needed for a Gibbs algorithm are $p(q_A|q_B, \theta_1, \theta_2, \mathcal{D}, \boldsymbol{y})$, $p(q_B|q_A, \theta_1, \theta_2, \mathcal{D}, \boldsymbol{y})$, $p(\theta_1|q_A, q_B, \theta_2\mathcal{D}, \boldsymbol{y})$, $p(\theta_2|q_A, q_B, \theta_1, \mathcal{D}, \boldsymbol{y})$ and the entire set $\{p(D_i|q_A, q_B, \theta_1, \theta_2, \mathcal{D}_{-i}, \boldsymbol{y}) : i = 1, \ldots, S\}$. At each cycle of the

Gibbs algorithm, then, there will be a need to sample from $4 + S$ distributions (as opposed to only 4 if the data model is available in closed form), where S is the number of sampling stations which is in the hundreds. Nevertheless, this may be the only option if the integral of expression (13) is not tractable or may be chosen anyway if the conditional posteriors of expression (24) are exceedingly difficult to sample from. Typically, although many more distributions are needed in this situation, the un-normalized conditional posterior densities allow substantial simplification.

Consider the example model of expressions (19) and (20) in which the D_i; $i = 1, \ldots, S$ are taken to be independent and identically distributed as gamma random variables. The joint posterior may then be represented as,

$$p(q_A, q_B, \alpha, \beta, \mathcal{D}, \boldsymbol{y}) \propto \prod_{i=1}^{S} \left[f(y_{A,i}, y_{B,i} | q_A, q_B, d_i)\, g(d_i|\alpha, \beta) \right] \pi_A(q_A)\, \pi_B(q_B)\, \pi_\alpha(\alpha)\, \pi_\beta(\beta) \tag{25}$$

Now, considerable simplification occurs as expression (25) is considered proportional to the conditional posteriors of q_A, q_B, α, β, and the elements of $\mathcal{D}$. In particular,

$$\begin{aligned}
p(q_A|q_B, \alpha, \beta, \mathcal{D}, \boldsymbol{y}) &\propto \prod_{i=1}^{S} \left[f(y_{A,i}, y_{B,i} | q_A, q_B, d_i) \right] \pi_A(q_A) \\
p(q_B|q_A, \alpha, \beta, \mathcal{D}, \boldsymbol{y}) &\propto \prod_{i=1}^{S} \left[f(y_{A,i}, y_{B,i} | q_A, q_B, d_i) \right] \pi_B(q_B) \\
p(\alpha|q_A, q_B, \beta, \mathcal{D}, \boldsymbol{y}) &\propto \prod_{i=1}^{S} \left[g(d_i|\alpha, \beta) \right] \pi_\alpha(\alpha) \\
p(\beta|q_A, q_B, \alpha, \mathcal{D}, \boldsymbol{y}) &\propto \prod_{i=1}^{S} \left[g(d_i|\alpha, \beta) \right] \pi_\beta(\beta) \\
\text{and, for } i = 1, \ldots, S & \\
p(d_i|q_A, q_B, \alpha, \beta, \mathcal{D}_{-i}, \boldsymbol{y}) &\propto f(y_{A,i}, y_{B,i} | q_A, q_B, d_i)\, g(d_i|\alpha, \beta)
\end{aligned} \tag{26}$$

These un-normalized density functions will be much simpler to simulate from than those corresponding to the closed form data model. For example, if $g(d_i|\alpha, \beta)$ is a gamma density and $\pi_\beta(\beta)$ is also a gamma density, then by conjugacy $p(\beta|q_A, q_B, \alpha, \mathcal{D}, \boldsymbol{y})$ will also be a gamma density, which may then be sampled from directly. This is exploitation of conditional conjugacy as mentioned in Section 3.4 on prior distributions. The last row of expression (26) gives S distributions, but each of these depends on only one bivariate pair of $(y_{A,i}, y_{B,i})$, not the entire data set. And, if $g(d_i|\alpha, \beta)$ is as given in expression (19) while $f(y_{A,i}, y_{B,i}|q_A, q_B)$ is as given in expression (20), then

$$p(d_i|q_A, q_B, \alpha, \beta, \mathcal{D}_{-i}, \boldsymbol{y}) \propto d_i^{\alpha+y_{A,i}+y_{B,i}-1} \exp\left\{-(q_A V_{A,i} + q_B V_{B,i} + \beta)d_i\right\},$$

so that each of these conditional posteriors are also identified as gamma distributions. Note that this great simplification is unlikely to occur if $g(d_i|\boldsymbol{\theta})$ must be chosen as a distribution with more extreme right tail than can be accommodated with a gamma distribution. Nevertheless, the point remains that although the Gibbs algorithm without a closed form data model requires simulation from many more conditional posterior distributions than does the algorithm with a closed form data model, the necessary sampling may be easy enough to make overall computation faster than use of the closed form – it will certainly make programming easier.

Simulation from the posterior of a beta-binomial model can be accomplished with a pure Metropolis-Hastings algorithm for jointly sampling the pair (μ, ϕ) defined at the end of the previous subsection. I have attached as an appendix to this report a lab exercise from the class Statistics 601: Advanced Methods at Iowa State that presents an example of analysis in considerable detail.

4.3 Posteriors for Calibration Factors

The procedures outlined in this section produce posterior distributions for any number of quantities, the primary interest being in the "catchabilities" q_A and q_B. Accounting for the volume (or, if appropriate for some species, area) towed complicates calculation of conversion factors to some degree, and one must consider what quantities are the object of conversion. To convert the catch on one tow from the Albatross to the corresponding value for the Bigelow under Model 1, *in a situation where each vessel conducts a tow* we would use the following.

$$\begin{aligned} E\{Y_{A,i}\} &= q_A V_{A,i} D_i \\ E\{Y_{B,i}\} &= q_B V_{B,i} D_i \end{aligned}$$

So then,

$$E\{Y_{B,i}\} = E\{Y_{A,i}\}\frac{E\{Y_{B,i}\}}{E\{Y_{A,i}\}} = E\{Y_{A,i}\}\frac{q_B V_{B,i} D_i}{q_A V_{A,i} D_i} = E\{Y_{A,i}\}\frac{q_B V_{B,i}}{q_A V_{A,i}},$$

so that the appropriate conversion factor is $\rho = (q_B V_{B,i}/q_A V_{A,i})$ rather than simply (q_B/q_A). Similarly, the appropriate conversion factor for Bigelow to Albatross will be $\phi = (q_A V_{A,i}/q_B V_{B,i})$. This presents one possible scenario for simulating posterior distributions of conversion factors, in which a posterior distribution of conversion factors exists *for each paired tow.* Assuming that a sufficient number of Monte Carlo iterations have been conducted for the sampling chains to have "converged" to the appropriate target distributions (i.e., burn-in) , simulated values of both conversion factors on a per-tow basis can be produced as follows. Let $q_{A,m}$ denote the value of q_A at the m^{th} cycle of the sampling algorithm and similarly for $q_{B,m}$. Then a value simulated from the posterior distribution of the conversion factor for converting expected Albatross catch to expected Bigelow catch for sampling station $i = 1, \ldots, S$ is,

$$\rho_{i,m} = \frac{q_{B,m} V_{B,i}}{q_{A,m} V_{A,i}}, \tag{27}$$

and a value simulated from the posterior distribution of the conversion factor for converting expected Bigelow catch to expected Albatross catch for sampling station $i = 1, \ldots, S$ is,

$$\phi_{i,m} = \frac{q_{A,m} V_{A,i}}{q_{B,m} V_{B,i}}. \tag{28}$$

Thus, each sampling station (i.e., each pair of tows) has its own posterior distribution for the conversion factors. These posterior distributions could be used for model assessment in determining how well we can predict the catch of one vessel from that of the other *for the data set in which both are available.* But for the objective of converting Bigelow to Albatross or Albatross to Bigelow for surveys conducted in either the past or the future this is not adequate because $V_{A,i}$ and $V_{B,i}$ are not both available.

Under the model, the catchabilities q_A and q_B are constant for a given situation (e.g., species or species group or species/season or species/sex combination). Despite this, when only one of the two vessels actually conducts a tow at a sampling station, and the desire is to convert to the corresponding value that would have occurred if the other vessel had conducted the tow instead, there are several ways to define what is desired in terms of a conversion factor.

1. One possible definition of conversion with data from a single vessel would be to predict the catch that would result from the other vessel if it had conducted a tow with the same volume (or area) sampled by the vessel actually conducting the tow. For example, to convert from Albatross to Bigelow in a situation in which only the Albatross was active (e.g., past surveys) assuming that the Bigelow sampled the same volume of water as the Albatross we would use

$$E\{\tilde{Y}_{B,i}\} = q_B V_{A,i} D_i = \frac{q_B}{q_A} q_A V_{A,i} D_i = \frac{q_B}{q_A} E\{Y_{A,i}\}, \tag{29}$$

   where in (29) $\tilde{Y}_{B,i}$ is used to denote the fact that this expected value differs from what we have taken as $E\{Y_{B,i}\}$ previously. Similarly, the conversion factor would be q_A/q_B for predicting the catch of the Albatross from that of the Bigelow, assuming that the Albatross had sampled the same volume or area as the Bigelow. This conversion is what occurs on an aggregate level (i.e., the total catch on a survey) if we simply multiple the total catch of a given vessel by the appropriate ratio of (estimated) catchabilities.

2. Another possible definition of conversion in this situation would necessitate

an assumption that $V_{A,i} = V_A$ and $V_{B,i} = V_B$ for all tows, and that these two constants are known. Under this assumption we may then consider predicting the catch of one vessel from that of the other if the vessel to be predicted had sampled using its own protocol (and, thus, its own constant volume or area). Then conversion is the same at the tow and aggregate levels. At the aggregate level, conversion from Albatross to Bigelow would be,

$$E\left\{\sum_{i=1}^{S} Y_{B,i}\right\} = q_B V_B \sum_{i=1}^{S} D_i = \frac{q_B V_B}{q_A V_A} q_A V_A \sum_{i=1}^{S} D_i = \frac{q_B V_B}{q_A V_A} \sum_{i=1}^{S} E\{Y_{A,i}\}. \qquad (30)$$

The appropriate factor for converting Albatross catch to Bigelow catch is then $\rho = (q_B V_B)/(q_A V_A)$ and that for converting Bigelow catch to Albatross catch is $\phi = (q_A V_A)/(q_B V_B)$.

Both of the above procedures rely on estimation of q_B/q_A or q_A/q_B. In the MCMC procedure outlined previously, samples from both of these quantities are automatically available. After convergence of the chain, at cycle $m = 1, \ldots, M$ of the MCMC algorithm we have values $q_{A,m}$ and $q_{B,m}$ available as sampled values from the respective posterior distributions. Thus, we also have available one value $\rho_m = (q_{B,m}/q_{A,m})$ sampled from the posterior distribution of this factor and one value $\phi_m = q_{A,m}/q_{B,m}$ sampled from the posterior distribution of this factor as well. With $m = 1, \ldots, M$ such simulated values an approximation to the posterior of each conversion factor is given by the empirical distribution of simulated values. Note that quantiles of these posteriors will be reciprocals and reversed. That is, the 25%$-$tile of the posterior distribution of $\rho = (q_B/q_A)$ will be one divided by the 75%$-$tile of the posterior distribution of $\phi = (q_A/q_B)$. It will not necessarily be the case, however, that expected values will be reciprocals.

The presentation of this subsection has been in the context of Model 1 but should apply equally to either Model 2 or Model 3. It will not be the case, however, for the beta-binomial model as has been mentioned previously. For this model, separate analyses will be needed for the two possible "directions" of conversion.

4.4 Model Assessment

For Model 1, Model 2, or Model 3 presented previously, assessment can be based on posterior predictive distributions, because these models include explicit distributional forms for random variables corresponding to observable quantities (e.g., $Y_{A,i}$ and $Y_{B,i}$). The posterior predictive distributions result from one additional step at each cycle of the MCMC algorithm. At iteration $m = 1, \ldots, M$ we have available sampled values $q_{A,m}$, $q_{B,m}$, $\boldsymbol{\theta}_m$, and $d_{i,m}$; $i = 1, \ldots, S$. For Model 1, an entire data set simulated from the posterior predictive distribution is then produced by additionally simulating values $y^*_{A,i}$ and $y^*_{B,i}$ from the data models $f(y_{A,i}|q_{A,m}, d_{i,m})$ and $f(y_{B,i}|q_{B,m}, d_{i,m})$. Assessment of model aptness based on posterior predictive distributions follows essentially the same progression as the model-based assessment described in Section 2. The major difference is simply how simulated data sets are produced – from a fitted model or a posterior predictive distribution.

Assessment of a beta-binomial model would require additional thought since this model conditions on the total catch at each sampling station. Comparison with the other models suggested here thus becomes more involved to avoid allowing this conditioning to *a priori* result in better agreement of the beta-binomial model with the data observed from the paired survey. As stated previously in Section 3.3.6, I find this formulation less than satisfying as a representation of the process by which the two vessels *catch fish*, although it may lead to quite good estimation of conversion factors *per se*. The use of cross-validation in conjunction with the posterior predictive distribution of conversion factors (as produced by a transformation of the binomial parameter for which we can easily simulate the posterior predictive) may be appropriate for this model. Setting aside 5% of the actual paired tows as a "prediction set", estimating the model based on the remaining data, and then predicting the catch of one vessel based only on the catch of the other (rather than the observed total) for the prediction set would remove the conditioning effect for the prediction data set. Repeating this for many randomly selected prediction sets could result in a powerful assessment for the beta-binomial formulation.

5 Borrowing Strength Within Species Groups

This section presents two versions of an overall procedure to meet the objective of producing stable estimates of conversion factors for situations (e.g., species or species/season or species/sex) for which only moderate amounts of data are available from the paired tow surveys. My personal preference lies with what is called "Bayes Empirical Bayes" in what follows, but I will also present what I will call "Hierarchical Extension" as an alternative, due to an objection that many statisticians would offer to Bayes Empirical Bayes. The objection, and my argument for preferring Bayes Empirical Bayes despite this objection, will be discussed in the last subsection of this section.

5.1 The Overall Intent

Both versions of the procedures outlined in what follows have the same objective and the same motivation, which draws directly on the discussion of Section 3.1 in this report. At the review, the panel spent considerable time pursuing the notion that any number of species have similar behaviors. Given that catchability is largely determined by the interaction of gear and sampling protocol with behavior, this suggests that catchabilities of the Albatross and Bigelow (q_A and q_B) may show distinct patterns across species groups. Also at the review, NMFS staff devoted considerable effort to the production of empirical evidence supporting or refuting this concept. I believe the result of these efforts was the demonstration that groups of similar species do, in fact, often have conversion factors with some commonality. For example, there appeared to be a distinctive “skate pattern” in relative catches of the two vessels and also a distinctive “flounder pattern”, among others. There were, to be certain, individual departures from generally perceived patterns, but the notion that interaction between sampling protocol (including gear) and fish behavior largely determines catchability seemed to be supported by empirical evidence from the available data. Staff from NMFS were readily able to begin development of a categorization of species groups based on factors such as whether species were demersal or pelagic, schooling or non-schooling, or exhibited what was called “herding behavior”. Calibration factors produced from the beta-binomial model by NMFS could be visually grouped along these same factors. This is an example of the combination of biological sensibility and observed data pattern mentioned in Section 3.1 of this report, and the objective is to use stochastic stability produced by these considerations to improve estimation of calibration factors for individual situations (e.g., species).

The fundamental conceptualization espoused in this section, then, is that biologically meaningful species groups result in distinct statistical distributions of conversion factors. Individual species differ in the appropriate conversion factor, but species within a group are more similar than species among groups. The basic intent is then to make use of information within groups of similar cases (i.e., species groups) to improve estimation in each individual case (i.e., species). As already mentioned, there are (at least) two approaches to accomplish this objective.

5.2 Bayes Empirical Bayes

I am uncertain whether or not the procedure outlined in this subsection has been called Bayes Empirical Bayes previously, but the name seems to fit. Any number of specific approaches to statistical analysis have been called Empirical Bayes, including what I would call Marginal Maximum Likelihood (e.g., maximization of a beta-binomial likelihood). One common formulation of what is sometimes called the "empirical Bayes problem" is that we are faced with a collection of like situations in each of which we wish to use the same form of likelihood and a common prior. A classic example is a collection of binomial problems. To select the common prior, we may estimate it through the use of a mixing distribution, such as a beta distribution in the example of a collection of binomial problems. This estimation is generally based only on the data, such as would occur with maximum likelihood. The mixing distribution with parameters set to the maximum likelihood estimates then becomes the common prior used for each individual problem in the collection. This is a traditional description of empirical Bayes. What is proposed here is similar, but also draws on the use of a posterior from "previous" analysis as the prior for a current situation.

Consider a collection of K species for which conversion factors are desired such that those species may be considered a species group with similar biological and behavioral characteristics. We wish to estimate the conversion factor separately for each species by producing posterior distributions for those factors, but would also like to account for commonality among these species by using the same *informative* prior. And, we would like that prior to be more strongly informative for individual species on which lesser data are available than for species on which greater data are available. To produce such a prior we consider the entire collection of data for the species group and produce a posterior using "naive" priors such as those described in Section 3.4 of this report. The basic idea is to use that posterior as the common prior for additional analyses that make use of only data from each individual species. There are two complications that need to be addressed to put this notion into practice. First, if the posterior from the combined analysis is produced through simulation, as will be the case here, no analytical parametric expression will be immediately available for its distributional form. Second, the strength of the combined posterior needs to be adjusted in a differential manner for each individual analysis depending on the amount of data available in those analyses.

1. Determining a Distributional Form.

   The more easily addressed of the two complications is turning an empirical posterior distribution of MCMC output into a analytical prior parametric distribution for use in individual analyses. This really amounts to nothing more than fitting a model to simulated values from a Markov chain; the simulated values are treated as "data". Such a fitting process need not be elaborate, as we are not fitting a distribution for the purposes of inference, and so care less about efficiency than bias, or at least consistency. Suppose, for example, we have determined that a gamma distribution is an appropriate mixing model for fish densities in Model 1 (i.e., the $g(d_i|\alpha, \beta)$ in that model, see expressions (19) and (20)). Suppose further that we have assigned q_A and q_B uniform priors on the interval (0, 1) as in expression (23), and have also assigned β a diffuse gamma prior and α an improper prior on the positive line as suggested in Section 3.4 of this report. We might then produce priors for analyses of individual species as follows.

   (a) Fit beta distributions using method of moments estimates to the simulated values of posterior distributions for both q_A and q_B.

   (b) Fit a gamma distribution also using the method of moments to the simulated values of the posterior distribution of β.

   (c) Fit an appropriate distribution, such as normal or gamma to the simulated values of the posterior distribution of α, depending on whether the empirical distribution of those simulated values appears more symmetric or right skew. If, by chance, the empirical distribution of these simulated values appears skew left (on the positive line) then one might fit an extreme value distribution. All of this can again be accomplished through simple matching of moments.

2. Discounting in Conversion of Posterior to Prior.

   The second difficulty noted previously was that we will desire a procedure to vary the strength of the prior distributions chosen for analyses of individual species. This may be accomplished through the process of "discounting" which refers simply to increasing the variance of the (modeled as in step 1) posterior from the combined analysis to arrive at a prior for analyses of individual species. It is difficult to avoid a certain amount of arbitrary decision making in this procedure, but the fundamental idea is to increase discounting as proportional to the amount of data (or a power of the amount of data) available for individual analyses. The exact mathematics of such a discounting process will depend on the particular forms chosen for parametric description of the posterior simulations from the combined analysis which, in turn, will depend on the particular forms chosen for the mixing and prior distributions of the model used in the combined analysis. In reports on a different project of estimation of discard in Northeast groundfish fisheries I have given a bit more specific suggestions for models particular to that problem. Similar ideas should apply here, although the exact procedure that might be best suited to the calibration problem will require additional thought.

5.3 Hierarchical Extension

An immediate objection to the Bayes Empirical Bayes procedure described in the previous subsection is that it makes double use of available data, once in determination of the prior (posterior from the combined analysis) and once in determination of the posterior (individual analyses). An alternative approach is based on the idea that considering individual species within a species group to be more similar that species from different groups motivates an additional level to the hierarchical models proposed previously. In this approach, we would consider catchabilities for individual species, $q_{A,k}$ and $q_{B,k}$ for species k say, to arise from an additional level in the hierarchy and be assigned either separate or joint mixing distributions $m_A(q_{A,k}|\psi_A)$ and $m_B(q_{B,k}|\psi_B)$ or $m(q_{A,k}, q_{B,k}|\psi)$. The catchabilities $q_{A,k}$ and $q_{B,k}$ are then also considered parameters in a simulation-based analysis (similar to the use made of the fish densities D_i in Section 4.2) and posteriors for these quantities produced as part of an overall MCMC simulation involving data from all of the members of a species group. This avoids the need for a "two-part" procedure consisting of an analysis for the combined data followed by an individual analysis for a particular species. It also avoids the double use of data contained in the Bayes Empirical Bayes approach. Complications introduced by this approach include the need to specify mixing distribution for the $q_{A,k}$ and $q_{B,k}$ and prior distributions for the parameters of those mixing distributions. Also, the complexity of the model would increase in at least the Poisson-based models because separate gamma mixing distributions for fish density would be needed for different species. As an example of this approach, consider Model 1 of Section 3.3.1 with a gamma chosen for the mixing distribution of fish densities across sampling stations as in Section 3.3.7. To extend this particular model to multiple species in a hierarchical manner we could proceed as follows. Let A and B denote Albatross and Bigelow and $i = 1, \ldots, S$ index sampling stations as before. Let $k = 1, \ldots, G$ index species within the species group of concern. Specify the following distributions.

$$\begin{aligned}
Y_{A,i,k} &\sim Po(q_{A,k}V_{A,i}D_{k,i}) \\
Y_{B,i,k} &\sim Po(q_{B,k}V_{B,i}D_{k,i}) \\
D_{k,i} &\sim Gamma(\alpha_k, \beta_k); \quad k = 1, \ldots, G
\end{aligned}$$

$$\begin{aligned} q_{A,k} &\sim Beta(a_k,\, b_k) \\ q_{B,k} &\sim Beta(a^*_k,\, b^*_k) \end{aligned}$$

Finally, priors would be needed on $\{(\alpha_k, \beta_k) : k = 1, \ldots, G\}$, (a_k, b_k), and (a^*_k, b^*_k), which could be achieved much as described in Section 3.4.

5.4 In Support of Bayes Empirical Bayes

As mentioned previously, at this point I prefer the Bayes Empirical Bayes approach to that of Hierarchical Extension. The reasons for this assessment follow.

1. The Bayes Empirical Bayes approach of fitting a model to the combined data from a species group and then using the posterior from that analysis to produce an (estimated) prior for analyses of individual species offers more control over prior dispersion than does an extension of the hierarchy. Experience gained in consideration of estimating amount of discard in another project suggests that situations involving large variability and small sample sizes are best approached through the use of strong prior information, based on either accumulated past data or biological knowledge. At the same time, situations in which a great deal more current data are available should not be overly constrained by prior strength. A hierarchical approach attempts to automate the degree to which individual situations borrow strength from the entire group. Although introducing a potentially arbitrary decision about how much discounting should be used in producing individual prior distributions from the group posterior, the Bayes Empirical Bayes approach allows greater flexibility in determining the degree to which strength is borrowed across individual situations.

2. The primary effect of the double use of data, as occurs in the Bayes Empirical Bayes procedure, is under-estimation of variability. That is, the prior is

selected to be “in concert” with the data and, as a result, the data tend to agree with the prior, giving a somewhat false sense of precision in the resultant posteriors. I believe that this can be admitted as a shortcoming of the Bayes Empirical Bayes approach without invalidating the approach in total. One could drop each individual species from the “group” model one-at-a-time to produce priors for individual analyses and this would eliminate the double use of data. My opinion is, however, that the increase in computational burden would outweigh an potential benefits.

6 Considerations of Length

Subsequent to the review meeting, I was asked to address the issue of how length might be incorporated into calibration models. There are two major avenues by which to accomplish incorporation of length if it is deemed important in a given situation, using length as a grouping factor, or using length as a continuous model covariate. It is not clear that greater ability of one vessel to capture smaller (or larger) fish should automatically result in length being incorporated as a continuous covariate in calibration models. A change in cod-end mesh size alone, for example, might well produce differential catchabilities for size classes, but would not necessarily indicate a continuous effect over an entire range of lengths. A difference in turbulence, on the other hand, might more plausibly be thought to have a continuous-like effect on the escape ability of fish of differing sizes. A combination of biological knowledge and empirical evidence would again seem like the best approach for determining whether length is important for a given situation and, if so, whether it should be considered a grouping factor or a continuous covariate.

6.1 Length as a Grouping Factor

If length is used as an additional grouping factor (e.g., small versus large) along with species and possibly season, then nothing new is introduced for analysis. The primary effect would be to create more individual situations with smaller sample sizes. There would be an additional question in the formation of groups for the approach of the previous section, however. In consideration of multiple species with multiple length classes, for example, should a combined group consist of a number of species all of the same length class, or of a number of species crossed with several length classes? In addition, unless a demarcation can be clearly identified from observed data, there may well be a certain arbitrariness to determining cut points for the formation of length classes.

6.2 Length as a Covariate

If length is used as a continuous covariate in a regression model, there is a need to determine the most appropriate way to incorporate it into a model. This will, to a large extent, be model specific, but a few general guidelines are possible. We consider separately here the Poisson-based models introduced in this report and the beta-binomial model already considered by NMFS.

6.2.1 Poisson Models

In the Poisson-based models of Section 3.3 (Models 1, 2, and 3) the most natural manner to incorporate a length effect would be through the conditional expected values $E\{Y_{A,i}|D_{A,i}\}$ and $E\{Y_{B,i}|D_{B,i}\}$. Under the data model of expression (11) in Section 3.2 these conditional expectations are,

$$\begin{aligned} E\{Y_{A,i}|D_{A,i}\} &= q_A V_{A,i} D_{A,i} \\ E\{Y_{B,i}|D_{B,i}\} &= q_B V_{B,i} D_{B,i} \end{aligned}$$

Now, however, we wish to split $Y_{A,i}$ and $Y_{B,i}$ into vectors of counts for different length classes. Let $j = 1, \ldots, \ell$ denote length classes to be used. Then the random variables become

$$\begin{aligned} Y_{A,i} &= (Y_{A,i,1}, Y_{A,i,2}, \ldots, Y_{A,i,\ell})^T \\ Y_{B,i} &= (Y_{B,i,1}, Y_{B,i,2}, \ldots, Y_{B,i,\ell})^T \end{aligned}$$

and the data models of expression (11) are replaced with, for $i = 1, \ldots, S$ and $j = 1, \ldots, \ell$,

$$\begin{aligned} f(y_{A,i,j}|D_{A,i,j}, q_{A,j}) &= \frac{1}{(y_{A,i,j})!}\{q_{A,j}V_{A,i}D_{A,i,j}\}^{y_{A,i,j}} \exp\{-q_{A,j}V_{A,i}D_{A,i,j}\} \\ f(y_{B,i,j}|D_{B,i,j}, q_{B,j}) &= \frac{1}{(y_{B,i,j})!}\{q_{B,j}V_{B,i}D_{B,i,j}\}^{y_{B,i,j}} \exp\{-q_{B,j}V_{B,i}D_{B,i,j}\} \end{aligned} \tag{31}$$

Note that, in (31), $q_{A,j}$ and $q_{B,j}$ are catchabilities for the j^{th} length class, the encountered fish densities are also indexed by length class, but the tow volumes (or areas) do not depend on length. The conditional expected values corresponding to the data model of (31) are now

$$\begin{aligned} E\{Y_{A,i,j}|D_{A,i,j}\} &= q_{A,j}V_{A,i}D_{A,i,j} \\ E\{Y_{B,i,j}|D_{B,i,j}\} &= q_{B,j}V_{B,i}D_{B,i,j} \end{aligned}$$

To model these conditional expectations so that catchabilities $q_{A,j}$ and $q_{B,j}$ are affected by length we can use a combination of several link functions from basic generalized linear models and what is typically called an “offset” term, this latter to separate $V_{A,i}$ and $D_{A,i,j}$ or $V_{B,i}$ and $D_{B,i,j}$ from the effects of length. The offset is the most easily understood and results from parameterization of a Poisson distribution in terms of a “rate”. Before proceeding to the actual data models (31) for this problem, consider the simpler case of Poisson responses for $Y_1, \ldots, Y_n$ such that, for a known constant w_i, Y_i has probability mass function

$$f(y_i|\rho_i, w_i) = \frac{1}{y_i!}(\rho_i w_i)^{y_i} \exp(-\rho_i w_i); \quad y_i = 0, 1, \ldots \tag{32}$$

Here, $E(Y_i) = \rho_i w_i$ and the interpretation is that ρ_i represents a "rate" for the occurrence of some event with a "population at risk" of size w_i. Note that defining ρ_i as a rate in this type of formulation requires a standard unit of "population" size, such as $100,000$ individuals. The parameters ρ_i are then expressed as number per unit and w_i is similarly expressed in terms of these units. While this formulation would be possible in principle for the current problem, determining a suitable unit of fish population would be problematic, particularly for multiple species. This difficulty aside, taking probability mass functions as in expression (32), a basic generalized linear model to incorporate covariates $\{x_i : i = 1, \ldots, n\}$ into the rate parameter ρ_i might use a log link function as,

$$\log\{E(Y_i)\} = \log(\rho_i) + \log(w_i) = \beta_0 + \log(w_i) + \beta_1 x_i \tag{33}$$

In (33) $\log(w_i)$ is called an "offset" term and notice that the covariate then effects only the rate parameter as $\log(\rho_i) = \beta_0 + \beta_1 x_i$.

We could use the device of an offset term with the data models in expression (31) which, with i indexing sampling station, j indexing length class, and x_j being the value of the j^{th} length class, would result in

$$\begin{aligned}
\log\{E(Y_{A,i,j}|D_{A,i,j})\} &= \log(q_{A,j}) + \log(V_{A,i}) + \log(D_{A,i,j}) \\
&= \beta_{A,0} + \log(V_{A,i}) + \log(D_{A,i,j}) + \beta_{A,1} x_j \\
\log\{E(Y_{B,i,j}|D_{B,i,j})\} &= \log(q_{B,j}) + \log(V_{B,i}) + \log(D_{B,i,j}) \\
&= \beta_{B,0} + \log(V_{B,i}) + \log(D_{B,i,j}) + \beta_{B,1} x_j
\end{aligned} \tag{34}$$

The only real elaboration of (34) over (33) is that the offset term has been split into two portions $\log(V_{A,i})$ and $\log(D_{A,i,j})$. Of course the settings differ in that

(33) was formulated for a simple problem involving one vector of response variables Y_i; $i = 1, \ldots, n$ while (34) contains multiple vectors (one for each sampling station indexed by i). A potential difficulty with (34) is that it implies models for the $q_{A,j}$ and $q_{B,j}$ as,

$$\begin{aligned} \log(q_{A,j}) &= \beta_{A,0} + \beta_{A,1} x_j \\ \log(q_{B,j}) &= \beta_{B,0} + \beta_{B,1} x_j \qquad (35) \end{aligned}$$

While this is analogous to the model for the ρ_i implied by (33) it could cause difficulties because, unlike the ρ_i in (33) for most problems, $q_{A,j}$ and $q_{B,j}$ have, for $j = 1, \ldots, \ell$, parameter spaces given by the unit interval (0, 1). Given that length class values x_j; $j = 1, \ldots, \ell$ will be strictly non-negative, one could ensure the proper parameter spaces by constraining $\beta_{A,0}$, $\beta_{B,0}$, $\beta_{A,1}$ and $\beta_{B,1}$ all to take values on the negative line. What seems a superior alternative in that it avoids the need for constraints, would be to model the $q_{A,j}$ and $q_{B,j}$ using a logit link function as

$$\begin{aligned} \log\left(\frac{q_{A,j}}{1 - q_{A,j}}\right) &= \beta_{A,0} + \beta_{A,1} x_j \\ \log\left(\frac{q_{B,j}}{1 - q_{B,j}}\right) &= \beta_{B,0} + \beta_{B,1} x_j \end{aligned}$$

or, with $\eta_{A,j} = \beta_{A,0} + \beta_{A,1} x_j$ and similarly for $\eta_{B,j}$,

$$\begin{aligned} q_{A,j} &= \frac{\exp(\beta_{A,0} + \beta_{A,1} x_j)}{1 + \exp(\beta_{A,0} + \beta_{A,1} x_j)} = \frac{\exp(\eta_{A,j})}{1 + \exp(\eta_{A,j})} \\ q_{B,j} &= \frac{\exp(\beta_{B,0} + \beta_{B,1} x_j)}{1 + \exp(\beta_{B,0} + \beta_{B,1} x_j)} = \frac{\exp(\eta_{B,j})}{1 + \exp(\eta_{B,j})} \qquad (36) \end{aligned}$$

From expression (36) we have that

$$\begin{aligned} \log(q_{A,j}) &= \eta_{A,j} - \log\{1 + \exp(\eta_{A,j})\} \\ \log(q_{B,j}) &= \eta_{B,j} - \log\{1 + \exp(\eta_{B,j})\} \qquad (37) \end{aligned}$$

Finally, substitution of (37) into the first lines of the expressions in (34) yields the models

$$\begin{aligned}
\log\{E(Y_{A,i,j}|D_{A,i,j})\} &= \log(q_{A,j}) + \log(V_{A,i}) + \log(D_{A,i,j}) \\
&= \eta_{A,j} - \log\{1 + \exp(\eta_{A,j})\} + \log(V_{A,i}) + \log(D_{A,i,j}) \\
\log\{E(Y_{B,i,j}|D_{B,i,j})\} &= \log(q_{B,j}) + \log(V_{B,i}) + \log(D_{B,i,j}) \\
&= \eta_{B,j} - \log\{1 + \exp(\eta_{B,j})\} + \log(V_{B,i}) + \log(D_{B,i,j})
\end{aligned} \tag{38}$$

In Section 3.3, three distinct models resulted from different assumptions about the distribution of fish densities encountered, which are now indexed by length class as $D_{A,i,j}$ and $D_{B,i,j}$. The same three assumptions may be applied to the data models of expression (31), resulting in analogous models to those presented previously, except taking length into consideration in each case. We might call these Model 1L, Model 2L, and Model 3L, respectively, with L denoting the incorporation of length in each model. Because the number of model "pieces" may have seemed to explode in consideration of length-based models, I give the length-based version of Model 1 (say Model 1L) explicitly.

Example: Model 1L

Model 1 of Section 3.3.1 arose through the assumption that $D_{A,i} = D_{B,i} = D_i$ in the data models of expression (11) and that $D_i \sim iid$ according to some distribution with density $g(d_i|\boldsymbol{\theta})$. The analogous assumption here is that $D_{A,i,j} = D_{B,i,j} = D_{i,j}$ in the data models of expression (31) for each $j = 1, \ldots, \ell$. The data models of expression (31) then become,

$$f(y_{A,i,j}|D_{i,j}, q_{A,j}) = \frac{1}{(y_{A,i,j})!}\{q_{A,j}V_{A,i}D_{i,j}\}^{y_{A,i,j}} \exp\{-q_{A,j}V_{A,i}D_{i,j}\}$$

$$f(y_{B,i,j}|D_{i,j}, q_{B,j}) \;=\; \frac{1}{(y_{B,i,j})!}\,\{q_{B,j}V_{B,i}D_{i,j}\}^{y_{B,i,j}}\,\exp\{-q_{B,j}V_{B,i}D_{i,j}\} \qquad (39)$$

where, as in expression (37) with x_j; $j = 1, \ldots, \ell$ denoting the value of length class j and $\eta_{A,j} = \beta_{A,0} + \beta_{A,1}x_j$ and $\eta_{B,j} = \beta_{B,0} + \beta_{B,1}x_j$,

$$\begin{aligned} \log(q_{A,j}) &= \eta_{A,j} - \log\{1 + \exp(\eta_{A,j})\} \\ \log(q_{B,j}) &= \eta_{B,j} - \log\{1 + \exp(\eta_{B,j})\} \end{aligned} \qquad (40)$$

Note that we could use any equivalent form here, such as the logistic equations given just before expression (37), or expression (37) itself.

Fish densities $D_{i,j}$ are taken to be independent and identically distributed across sampling stations i and independent among length classes j, so that we me write the mixing distributions for $D_{i,j}$ as $g(d_{i,j}|\boldsymbol{\theta}_j)$ for $j = 1, \ldots, \ell$. If, for example, these are taken as gamma distributions, we would have $\boldsymbol{\theta}_j \equiv (\alpha_j,\, \beta_j)$ and,

$$g(d_{i,j}|\alpha_j, \beta_j) = \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)}\, d_{i,j}^{\alpha_j}\, \exp(-\beta_j d_{i,j}); \quad d_{i,j} > 0 \qquad (41)$$

For a Bayesian analysis of this model we need to assign prior distributions to the parameters $\beta_{A,0}$, $\beta_{B,0}$, $\beta_{A,1}$, $\beta_{B,1}$, and $\{(\alpha_j, \beta_j) : j = 1, \ldots, \ell\}$. A natural choice is to assign all of the regression parameters normal distributions,

$$\begin{aligned} \beta_{A,0} &\sim N(m_{A,0}, v_{A,0}) \\ \beta_{A,1} &\sim N(m_{A,1}, v_{A,1}) \\ \beta_{B,0} &\sim N(m_{B,0}, v_{B,0}) \\ \beta_{B,1} &\sim N(m_{B,1}, v_{B,1}) \end{aligned} \qquad (42)$$

In (42) a simple device would be to set $m_{A,0} = m_{B,0} = m_{A,1} = m_{B,1} = 0$, which would give a prior expectation for all $q_{A,j}$ and $q_{B,j}$ of 0.5. This could be easily adjusted by setting $m_{A,0} = m_{B,0} = c$ for some constant c; values of c greater than zero lead to prior expected values greater than 0.5 and negative values of c lead to prior expected values less than 0.5. The variances could be all taken to produce diffuse priors, for example $v_{A,0} = v_{B,0} = v_{A,1} = v_{B,1} = 100$ (or 1000). Prior distributions for the parameters $\{(\alpha_j, \beta_j) : j = 1, \ldots, \ell\}$ could be specified in the same manner as discussed in Section 3.4, with $\pi(\alpha_j, \beta_j) = \pi(\alpha_j)\pi(\beta_j)$ where $\pi(\beta_j)$ is diffuse gamma (e.g., gamma$(0.01, 0.01)$ or gamma$(0.001, 0.001)$) and $\pi(\alpha_j)$ is either improper on the positive line or proper uniform on a large interval.

Given specification of priors, full conditional posteriors needed for a Gibbs sampling algorithm similar to that described in Section 4.2 combine prior distributions and the data model. Let $\boldsymbol{Y}_A = \{Y_{A,i,j} : i = 1, \ldots, S; j = 1, \ldots, \ell\}$ and $\boldsymbol{Y}_B \equiv \{Y_{B,i,j} : i = 1, \ldots, S; j = 1, \ldots, \ell\}$ denote the entire collection of random variables across all sampling stations and length classes for the two vessels. Similarly, let $\boldsymbol{\mathcal{D}} \equiv \{D_{i,j} : i = 1, \ldots, S; j = 1, \ldots, \ell\}$. Also, let $\boldsymbol{\beta}_A \equiv (\beta_{A,0}, \beta_{A,1})^T$ and $\boldsymbol{\beta}_B \equiv (\beta_{B,0}, \beta_{B,1})^T$ denote the regression parameters for the two vessels. Using conditional independence between vessels, among length classes within each sampling station, and across sampling stations,

$$\begin{aligned} f(\boldsymbol{y}_A|\boldsymbol{\beta}_A, \mathcal{D}) &= \prod_{i=1}^{S}\prod_{j=1}^{\ell} f(y_{A,i,j}|\boldsymbol{\beta}_A, D_{i,j}) \\ f(\boldsymbol{y}_B|\boldsymbol{\beta}_B, \mathcal{D}) &= \prod_{i=1}^{S}\prod_{j=1}^{\ell} f(y_{B,i,j}|\boldsymbol{\beta}_B, \mathcal{D}_{i,j}) \end{aligned} \tag{43}$$

Note that, in (43), $f(y_{A,i,j}|\boldsymbol{\beta}_A, D_{i,j})$ denotes the data model (39) combined with the regression model (40), and similarly for $f(y_{B,i,j}|\boldsymbol{\beta}_B, D_{i,j})$. Use the notation $p(x|\cdot)$ to denote the posterior of x given "everything else", where $x \in \{\beta_{A,0}, \beta_{A,1}, \beta_{B,0}, \beta_{B,1}, \{D_{i,j} : i = 1, \ldots, S; j = 1, \ldots, \ell\}, \{\alpha_j : j = 1, \ldots, \ell\}, \{\beta_j : j = 1, \ldots, \ell\}\}$ and "everything else" is all members of this set other than x. Then the necessary full conditional posteriors for the regression parameters are,

$$\begin{aligned} p(\beta_{A,0}|\cdot) &\propto \pi(\beta_{A,0}) f(\boldsymbol{y}_A|\boldsymbol{\beta}_A, \mathcal{D}) \\ p(\beta_{A,1}|\cdot) &\propto \pi(\beta_{A,1}) f(\boldsymbol{y}_A|\boldsymbol{\beta}_A, \mathcal{D}) \end{aligned}$$

$$\begin{aligned} p(\beta_{B,0}|\cdot) &\propto \pi(\beta_{B,0}) f(\boldsymbol{y}_B|\boldsymbol{\beta}_B, \mathcal{D}) \\ p(\beta_{B,1}|\cdot) &\propto \pi(\beta_{B,1}) f(\boldsymbol{y}_B|\boldsymbol{\beta}_B, \mathcal{D}) \end{aligned} \qquad (44)$$

Those for the controlling parameters of the fish density mixing distributions are, for $j = 1, \ldots, \ell$

$$\begin{aligned} p(\alpha_j|\cdot) &\propto \pi(\alpha_j) \prod_{i=1}^{S} g(d_{i,j}|\alpha_j, \beta_j) \\ p(\beta_j|\cdot) &\propto \pi(\beta_j) \prod_{i=1}^{S} g(d_{i,j}|\alpha_j, \beta_j) \end{aligned} \qquad (45)$$

And those for the latent (auxiliary) densities of fish are, for $i = 1, \ldots, S$ and $j = 1, \ldots, \ell$,

$$p(d_{i,j}|\cdot) \propto g(d_{i,j}|\alpha_j, \beta_j) f(y_{A,i,j}|\boldsymbol{\beta}_A, d_{i,j}) f(y_{B,i,j}|\boldsymbol{\beta}_B, d_{i,j}) \qquad (46)$$

A Gibbs sampling algorithm would then take the form described in Section 4.2, using the marginal data model as represented by unevaluated integrals. The production of posteriors for length-based calibration factors should be a straightforward extension of the material presented in Section 4.3, and the same for model assessment as described in Section 4.4. Use of models that incorporate length in a procedure that borrows strength within a group of species would almost necessarily rely on the Bayes Empirical Bayes approach of Section 5.2 to avoid potential problems with model identifiability in the normal priors for regression parameters.

6.2.2 Beta-Binomial Formulation

It is also possible to incorporate length as a covariate in the beta-binomial model presented by NMFS at the review meeting. Before presenting this model formulation, however, it is worth noting that one could also choose to incorporate length as a covariate directly into a binomial data model, and then let the regression coefficients be random across sampling sites. I believe this model structure has less justification than that presented in what follows because it would imply that the *relation* between length and proportion of total catch at each sampling site due to the Bigelow (or Albatross) varies across sampling sites. If the relation between length of fish and characteristics of gear and sampling protocol of the two vessels is due to interaction between fish behavior and gear/towing configuration, this would seem an untenable assumption. The proportion of total catch accounted for by one vessel might vary due to the effects of fish density, but the relation between this and length, if there is one, should be characteristic of vessel gear and sampling protocol. Thus, although one could certainly formulate a model in which the relation between length and proportion of total catch due to one vessel varies across sampling sites, my opinion is that the model presented here is more defensible.

To formulate the beta binomial model in a manner appropriate for incorporation of length, let $Z_{i,j}$ be a random variable connected with the catch of the Bigelow for sampling station i and length class j, conditioned on the total catch of both vessels having a value of $c_{i,j}$. The possible values of $Z_{i,j}$ are then contained in the set $\{0, 1, \ldots, c_{i,j}\}$. The data model is then,

$$f(z_{i,j}|\theta_{i,j}) = \frac{c_{i,j}!}{z_{i,j}!\,(c_{i,j} - z_{i,j})!}\, z_{i,j}^{\theta_{i,j}}\,(1-\theta_{i,j})^{c_{i,j}-z_{i,j}}; \quad z_{i,j} = 0, 1, \ldots, c_{i,j} \tag{47}$$

Now, let the $\theta_{i,j};\ i = 1, \ldots S;\ j = 1, \ldots, \ell$ be independent with mixing distributions

$$g(\theta_{i,j}|\alpha_j, \beta_j) = \frac{\Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)}\, \theta_{i,j}^{\alpha_j - 1}\,(1-\theta_{i,j})^{\beta_j - 1}; \quad 0 < \theta_{i,j} < 1 \tag{48}$$

The expected value and variance of the distribution in expression (48) are

$$\begin{aligned} E(\theta_{i,j}) &= \frac{\alpha_j}{\alpha_j + \beta_j} = \mu_j \\ var(\theta_{i,j}) &= \frac{\alpha_j\,\beta_j}{(\alpha_j + \beta_j)^2\,(\alpha_j + \beta_j + 1)} = \phi\,\mu_j(1-\mu_j) \end{aligned} \tag{49}$$

where $\phi = 1/(\alpha_j + \beta_j)$.

Note that we have imposed a restriction in expression (49). The expected values of the $\theta_{i,j}$ are allowed to vary with length class $j = 1, \ldots, \ell$, but the original beta distribution parameters are constrained in that $\alpha_j + \beta_j$ is constant across length classes. Nevertheless, parameterization in terms of μ_j and ϕ has served to isolate expected values in terms of the parameters μ_j; $j = 1, \ldots, \ell$. Notice also that the relevant parameter spaces are now $0 < \mu_j < 1$ and $0 < \phi < 1$. This allows incorporation of length as the covariate x_j,

$$\log\left(\frac{\mu_j}{1-\mu_j}\right) = \beta_0 + \beta_1 x_j. \tag{50}$$

As a side note, other link functions could be used in place of the logit link in expression (50) if they were deemed more appropriate, such as a log-log link or a complimentary log-log link. To complete a Bayesian specification of this model requires prior distributions for β_0, β_1 and ϕ. The natural choices are

$$\begin{aligned} \beta_0 &\sim N(m_0, v_0) \\ \beta_1 &\sim N(m_1, v_1) \\ \phi &\sim U(0, 1) \end{aligned} \tag{51}$$

Setting $m_0 = m_1 = 0$ in (51) would lead to a prior expectation for μ_j of 0.50. The prior parameters v_0 and v_1 could again be made large, as suggested for the normal priors in expression (42) for Model 1L.

This beta-binomial model contains only three parameters, although if the $\{\theta_{i,j} : i = 1, \ldots, S;\ j = 1, \ldots, \ell\}$ are not integrated out of the data model, we will need $3 + S\ell$ conditional posteriors for simulation. Write the joint prior in product form $\pi(\beta_0, \beta_1, \phi) = \pi(\beta_0)\pi(\beta_1)\pi(\phi)$ where $\pi(\beta_0)$, $\pi(\beta_1)$ and $\pi(\phi)$ are as specified in (51). Let $g(\theta_{i,j}|\beta_0, \beta_1, \phi)$ denote the mixing distribution (48) written in terms of parameters β_0, β_1 and ψ, as results from the re-parameterization of (49) and the model of (50). Let $\boldsymbol{\theta} \equiv \{\theta_{i,j} : i = 1, \ldots, S;\ j = 1, \ldots \ell\}$ and $\boldsymbol{\theta}_{-(i,j)} \equiv \boldsymbol{\theta} \backslash \theta_{i,j}$ be this set less the value $\theta_{i,j}$. Then the necessary full conditional posteriors for the three fixed parameters may be written as

$$\begin{aligned} p(\beta_0|\beta_1, \phi, \boldsymbol{\theta}, \boldsymbol{z}) &\propto \pi(\beta_0) \prod_{i=1}^{S} \prod_{j=1}^{\ell} g(\theta_{i,j}|\beta_0, \beta_1, \phi) \\ p(\beta_1|\beta_0, \phi, \boldsymbol{\theta}, \boldsymbol{z}) &\propto \pi(\beta_1) \prod_{i=1}^{S} \prod_{j=1}^{\ell} g(\theta_{i,j}|\beta_0, \beta_1, \phi) \\ \pi(\phi|\beta_0, \beta_1, \boldsymbol{\theta}, \boldsymbol{z}) &\propto \pi(\phi) \prod_{i=1}^{S} \prod_{j=1}^{\ell} f(z_{i,j}|\theta_{i,j}) \end{aligned} \tag{52}$$

and those for the $\theta_{i,j};\ i = 1, \ldots, S;\ j = 1, \ldots, \ell$ as

$$p(\theta_{i,j}|\beta_0, \beta_1, \phi, \boldsymbol{\theta}_{-(i,j)}, \boldsymbol{z}) \propto g(\theta_{i,j}|\beta_0, \beta_1, \phi)\, f(z_{i,j}|\theta_{i,j}) \tag{53}$$

Because each of the $S\ell$ distributions in (53) will be the product of a beta density with a binomial probability mass function, each of these conditional posteriors will be beta distributions and simulation of the $\theta_{i,j}$ will be fast. An alternative would be to integrate the $\theta_{i,j}$ out of the data model as described in Section 4.1 but, given the simple form of the distributions in expression (53) this likely would not produce a substantial decrease in computational effort. The primary difficulty in this MCMC algorithm will be sampling the first two distributions in expression (52); the conditional posterior of ϕ will again be a beta distribution and will not cause any problems. As for the Poisson-based models of Section 6.2.1, use of a beta-binomial formulation including length as a covariate to borrow strength within a species group would be feasible perhaps only through the use of what has been called here Bayes Empirical Bayes. As with those previous models, attempts to extend the hierarchy with a beta-binomial formulation could easily result in non-identifiable models.

7 Summary

There is a fair amount of material contained in this report. This is due, in part, to the attempt to present a range of possible models within a consistent framework, rather than select one possibility and present it as the only alternative. I have also attempted to provide a moderate amount of detail for several cases to make the suggestions offered more concrete. Full implementation of any of the ideas offered in this report, however, will require additional effort, and NMFS faces some immediate needs related to stock assessments. At the same time, the procedures currently developed for estimation of calibration factors will not prove adequate for the full suite of situations in which such estimation is desired. My recommendations for making use of the ideas contained in this report follows.

1. The Chair's consensus report recommends a procedure to determine a methodology for estimation of calibration factors in pressing situations. I agree that this is a defensible strategy that should be followed in the near-term. The one addition to this prescription, that was also recommended by the review panel, is a more thorough assessment of representation of zero frequencies occurrences. Methods by which this can be accomplished in a non-Bayesian setting are presented in Section 2 of this report. The outcome of such an assessment should be a determination of whether the data models under current consideration, including the beta-binomial model and the Poisson-based models presented in this report, are adequate to reflect the frequencies of zero catches in the data. The alternative is to replace those data models with finite mixtures or so-called "zero-inflated" models. Note that everything presented in this report would apply to this type of data model as well as to the formulations that have been given explicitly, although care would be needed in derivation of exact mathematical expressions for zero-inflated models.

2. Assessment of the general modeling framework presented in Section 3 of this report should be pursued beginning immediately. This assessment concerns the use of Poisson-based models that take catchability of vessels, as determined by the interaction of fish behavior with gear and sampling protocols,

as fixed values. The central question is whether such a formulation of the problem (possibly with finite mixture or zero-inflated data models) should be preferred to the beta-binomial formulation that conditions on total catch by both vessels at each sampling station and takes proportional catch by one vessel (and, hence, catchability) as random across stations. Poisson-based formulations with catchability as characteristics of vessels seem preferable in terms of problem conceptualization. But, as mentioned in this report, that does not necessarily guarantee superior performance in estimation of calibration factors.

3. Incorporation of length as either a grouping factor or continuous covariate is conceptually straightforward with any of the Poisson-based models or the beta-binomial model. Mathematical details needed to result in practical estimation procedures, however, are not trivial and depend largely on re-parameterization of models to isolate expected values and attention to issues of consistency in allowable parameter spaces under various parameterizations of statistical distributions. Section 6 of this report contains guidance on how this can be accomplished using either Poisson-based or beta-binomial models.

4. The most statistically controversial suggestion in this report is the use of what has been called Bayes Empirical Bayes in Section 5 to allow estimation of calibration factors for situations with modest sample sizes. This procedure is not a part of standard statistical practice, although I find it hard to believe that no one has thought of it before and there may very well be some information about this approach in the literature. While I believe this is a practical and defensible procedure, it will most likely result in a certain degree of underestimation of uncertainty. Thus, although I recommend this method be used, I acknowledge the need for additional research into its properties.

Appendix: Bayesian Analysis of a Beta-Binomial Model

Statistics 601, Fall 2008

Lab 14 Notes

This lab concerns a Bayesian analysis of beta-binomial mixture models. We will consider both a simulated example and an analysis of the *Gambusia* reproduction study described in Example 7.12 of the course notes and continued as Example 9.2, this latter in which we looked at forming interval estimates using normed profile likelihoods. We approach this problem as a Bayesian analysis of a beta-binomial mixture model, rather than an analysis of a collection of binomial models to which we wish to assign a hierarchical prior.

Model:

Assume we have a collection of observable random variables $Y_1, \ldots, Y_n$ and fixed constants $m_1, \ldots, m_n$ such that the distributions of of the random variables are taken as independent for given values of parameters $\theta_1, \ldots, \theta_n$. Specifically, let Y_i have pmf

$$f(y_i|\theta_i) = \frac{m_i!}{(m_i - y_i)!\, y_i!}\, \theta_i^{y_i}\, (1 - \theta_i)^{m_i - y_i}; \quad y_i = 0, 1, \ldots, m_i \tag{54}$$

and with $0 < \theta_i < 1$ for $i = 1, \ldots, n$. Further, let θ_i be exchangeable random variables having pdf, for $i = 1, \ldots, n$,

$$g(\theta_i|\alpha,\, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\, \Gamma(\beta)}\, \theta_i^{\alpha - 1}\, (1 - \theta_i)^{\beta - 1}; \quad 0 < \theta_i < 1 \tag{55}$$

As we are considering this a Bayesian analysis of a beta-binomial mixture model, we have little direct interest in possible values of the θ_i; $i = 1, \ldots, n$ and it is appropriate to determine the distributions of the Y_i as they depend only on the parameters α and β. As we have seen previously, these are independent mixture distributions of the form,

$$\begin{aligned} h(y_i|\alpha,\, \beta) &= \int_0^1 f(y_i|\theta_i)\, g(\theta_i|\alpha,\, \beta)\, d\theta_i \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, \frac{\Gamma(\alpha+y_i)\,\Gamma(\beta+m_i-y_i)}{\Gamma(\alpha+\beta+m_i)}\, \frac{m_i!}{(m_i-y_i)!\,y_i!} \qquad (56) \end{aligned}$$

The joint mixture for $Y_1, \ldots, Y_n$ given parameters α and β is the product of the n pmfs in expression (3),

$$h(\boldsymbol{y}|\alpha,\, \beta) = \prod_{i=1}^{n} h(y_i|\alpha,\, \beta). \qquad (57)$$

To assign prior distributions to the parameters of the joint mixture model (4) it is convenient to first re-parameterize in terms of the mean of the beta mixing distribution and an additional parameter. The mean, which is $\mu = \alpha/(\alpha+\beta)$ has range (0, 1). In analysis by maximum likelihood, we took the additional parameter to be $\phi = 1/(\alpha+\beta)$, which has range $(0,\, \infty)$. While we could certainly work with this, it is even more convenient to choose $\phi = 1/(\alpha+\beta+1)$ which has range (0, 1) so that we have two parameters that both assume values on the unit interval. This will also make it easier to simulate values from the posterior $p(\alpha,\, \beta|\boldsymbol{y})$ (even if we don't arrive at a very efficient algorithm).

To accomplish the re-parameterization then, let

$$\begin{aligned} \mu &= \frac{\alpha}{\alpha+\beta} \\ \phi &= \frac{1}{\alpha+\beta+1}, \end{aligned}$$

so that,

$$\begin{aligned} \alpha &= \frac{(1-\phi)\,\mu}{\phi} \\ \beta &= \frac{(1-\mu)\,(1-\phi)}{\phi} \\ \alpha+\beta &= \frac{1-\phi}{\phi}. \qquad (58) \end{aligned}$$

The joint mixture model of expression (4) may then be written in terms of μ and ϕ as,

$$\begin{aligned} h(\boldsymbol{y}|\mu,\, \phi) &= \left(\prod_{i=1}^{n} \frac{m_i!}{(m_i - y_i)!\, y_i!}\right) \left[\frac{\Gamma\left(\frac{1-\phi)}{\phi}\right)}{\Gamma\left(\frac{(1-\phi)\,\mu}{\phi}\right)\, \Gamma\left(\frac{(1-\phi)\,(1-\mu)}{\phi}\right)}\right]^n \\ &\times \frac{\prod_{i=1}^{n} \Gamma\left(\frac{(1-\phi)\,\mu}{\phi} + y_i\right) \prod_{i=1}^{n} \Gamma\left(\frac{(1-\mu)(1-\phi)}{\phi} + m_i - y_i\right)}{\prod_{i=1}^{n} \Gamma\left(\frac{1-\phi}{\phi} + m_i\right)}. \qquad (59) \end{aligned}$$

Now, expression (6) looks fairly horrid and it, in fact, is. However, in computer functions we can always compute $h(\boldsymbol{y}|\mu,\, \phi)$ by expressions (3) and (4) after first defining α and β in terms of μ and ϕ as in expression (5).

Prior Distributions

Because we now have parameters μ and ϕ such that $\mu \in (0,\, 1)$, $\phi \in (0,\, 1)$ and $(\mu,\, \phi) \in (0,\, 1) \times (0,\, 1)$, we can assign a joint prior distribution as,

$$\pi(\alpha,\, \beta) = \pi(\alpha)\, \pi(\beta), \qquad (60)$$

where $\pi(\cdot)$ is a uniform distribution on the interval (0, 1). That is,

$$\pi(\mu) = \begin{cases} 1 & 0 < \mu < 1 \\ 0 & o.w. \end{cases}$$

and,

$$\pi(\phi) = \begin{cases} 1 & 0 < \phi < 1 \\ 0 & o.w. \end{cases}$$

Posterior Distribution

The joint posterior distribution of $(\mu,\, \phi)$ is such that,

$$p(\mu,\, \phi) \propto h(\boldsymbol{y}|\mu,\, \phi)\, \pi(\mu,\, \phi) = h(\boldsymbol{y}|\mu,\, \phi), \qquad (61)$$

where $h(\boldsymbol{y}|\mu,\, \phi)$ is given in (6) and $\pi(\mu,\, \phi)$ is given in (7), and this expression holds for any $(\mu,\, \phi) \in (0,\, 1) \times (0,\, 1)$.

This joint posterior will need to be assessed through simulation rather than analytical derivation. We have already seen the use of a Gibbs Sampling algorithm to simulate from a joint posterior, and we could certainly consider such an algorithm here, considering (8) first as a function of μ for a given ϕ and then again as a function of ϕ for a given μ. This would, however, perhaps be more cumbersome than needed, and we might like to simulate from a joint distribution with density proportional to (8) directly. An approach by which to accomplish this is provided by a Metropolis algorithm, which we described in a previous lab.

Simulation from Joint Posterior

To simulate from $p(\mu,\, \phi|\boldsymbol{y})$ using a Metropolis algorithm, we need (1) a candidate distribution by which to produce proposed "jumps" for the sampler and (2) calculation of the probability for accepting proposed jumps. The first of these is fairly easy in this problem because the joint "sample" space of $(\mu,\, \phi)$ is $(0,\, 1) \times (0,\, 1)$, which suggests an independence chain with candidate distribution

$$f(\mu,\, \phi) = \begin{cases} 1 & \text{if } (\mu,\, \phi) \in\ (0,\, 1) \times (0,\, 1) \\ 0 & \qquad\qquad o.w. \end{cases} \tag{62}$$

We may easily simulate values $(\mu^*,\, \phi^*)$ from this distribution by simulating μ^* and ϕ^* independently from uniform distributions on the unit interval.

The Metropolis acceptance probability for proposed jumps from a current value $(\mu,\, \phi)$ to a new proposed value $(\mu^*,\, \phi^*)$ takes the form of

$$\begin{aligned} \alpha'[(\mu,\, \phi),\, (\mu^*,\, \phi^*)] &= \min\{h(\boldsymbol{y}|\mu^*,\, \phi^*)/h(\boldsymbol{y}|\mu,\, \phi),\ 1\} \\ &= \min\{w(\mu^*,\, \phi^*,\, \mu,\, \phi),\ 1\}, \end{aligned} \tag{63}$$

where we have denoted this probability as α' so as not to confuse it with the parameter α in the mixture model.

Our essential difficulty at this point is computation of the ratio $w(\mu^*,\, \phi^*,\, \mu,\, \phi)$, which is complicated by the fact that $h(\boldsymbol{y}|\mu,\, \phi)$ in (6) contains ratios of products of gamma functions. Such functions can easily assume either huge or negligible values, resulting in computational values of infinity or values that fail to exist (i.e., the NaN assignment in R or Splus). As a result, even though the ratios may be well within normal “computational bounds”, the components in the numerator or denominator are not, producing computation algorithms to fail.

Our solution to this difficulty rests on two computational techniques that are both worth knowing. First, note that the form of $w(\cdot, \cdot)$ is a ratio, and any ratio may be written as the exponentiation of the difference of logarithms. Specifically,

$$\begin{aligned} w(\mu^*,\, \phi^*,\, \mu,\, \phi) &= \frac{h(\boldsymbol{y}|\mu^*,\, \phi^*)}{h(\boldsymbol{y}|\mu,\, \phi)} \\ \\ &= \exp\left[\log\{h(\boldsymbol{y}|\mu^*,\, \phi^*)\} - \log\{h(\boldsymbol{y}|\mu,\, \phi)\}\right]. \end{aligned} \tag{64}$$

Also note that, from (4) we have that,

$$\log\{h(\boldsymbol{y}|\mu,\, \phi)\} = \sum_{i=1}^{n} \log\{h(y_i|\mu,\, \phi)\}. \tag{65}$$

Secondly, notice that the components of (12) can be simplified by applying properties of gamma functions to the component densities in (3). That is, (4) may be written as proportional to,

$$h(\boldsymbol{y}|\alpha,\, \beta) \propto \prod_{i=1}^{n} \frac{\prod_{j=0}^{y_i-1}(\alpha+j)\, \prod_{j=0}^{m_i-y_i-1}(\beta+j)}{\prod_{j=0}^{m_i-1}(\alpha+\beta+j)}, \tag{66}$$

or, in terms of logs,

$$\log\{h(\boldsymbol{y}|\alpha,\, \beta)\} \propto \sum_{i=1}^{n} \left[\sum_{j=0}^{y_i-1} \log(\alpha+j) \quad \sum_{j=0}^{m_i-y_i-1} \log(\beta+j) - \sum_{j=0}^{m_i-1} (\alpha+\beta+j) \right], \tag{67}$$

and (14) can be computed in terms of μ and ϕ by first defining α and β as in expression (5).

All of this leads to a practical computational approach for implementation of a Metropolis algorithm to (1) generate candidate jumps and (2) calculate acceptance probabilities for those proposals. An outline of that algorithm is as follows.

1. Begin with initial values $(\mu_0, \phi_0) \in (0, 1) \times (0, 1)$
2. Set current values $(\mu_c, \phi_c) = (\mu_0, \phi_0)$ and set $k = 1$
3. At iteration k, generate a proposed jump (μ^*, ϕ^*) as a pair of independent values from uniform distributions on the unit interval.
4. Compute $w(\mu^*, \phi^*, \mu_c, \phi_c)$ as given in (11) making use of (14) with (α_c, β_c) and (α^*, β^*) defined by the transformations in (5).
5. Generate an independent value u from a uniform distribution on (0, 1).
6. If $u \leq w(\mu^*, \phi^*, \mu_c, \phi_c)$ let $(\mu_k, \phi_k) = (\mu^*, \phi^*)$. Otherwise, let $(\mu_k, \phi_k) = (\mu_c, \phi_c)$
7. Set $(\mu_c, \phi_c) = (\mu_k, \phi_k)$ and update k to $k + 1$, and return to step 3.
8. Discard values for a burn-in period $(k \leq B)$
9. Continue for M additional iterations, collecting values of (μ_k, ϕ_k) at each iteration.

At the conclusion of this algorithm we have a collection of M values of (μ, ϕ) simulated from the posterior $p(\mu, \phi|\boldsymbol{y})$.

Simulation from Posterior Predictive

An important part of model assessment in many Bayesian analyses is the ability to simulate not only from the joint (and hence marginal) posterior distribution(s), but also from the posterior predictive distribution. In a situation involving a data model $f(\boldsymbol{y}|\theta)$ and a prior $\pi(\theta)$, the posterior predictive distribution is defined as the distribution for a new observation (or new data set) conditional on the observed data. Let $\boldsymbol{y}^*$ denote a new (or future) observation assumed to follow the same data model as $\boldsymbol{y}$ so that, given θ, the distribution is $f(\boldsymbol{y}^*|\theta)$. The posterior predictive distribution is then,

$$\begin{aligned} p(\boldsymbol{y}^*|\boldsymbol{y}) &= \frac{h(\boldsymbol{y}^*,\, \boldsymbol{y})}{h(\boldsymbol{y})} \\ &= \frac{\int f(\boldsymbol{y}^*|\theta)\, f(\boldsymbol{y}|\theta)\, \pi(\theta)\, d\theta}{h(\boldsymbol{y})} \\ &= \int f(\boldsymbol{y}^*|\theta)\, \frac{f(\boldsymbol{y}|\theta)\, \pi(\theta)}{h(\boldsymbol{y})}\, d\theta \\ &= \int f(\boldsymbol{y}^*|\theta) p(\theta|\boldsymbol{y})\, d\theta \end{aligned} \tag{68}$$

This is the structure of the model we are using here, with $f(\boldsymbol{y}^*|\theta)$ and $f(\boldsymbol{y}|\theta)$ replaced by $h(\boldsymbol{y}^*|\mu,\, \phi)$ and $h(\boldsymbol{y}|\mu,\, \phi)$, and $\pi(\theta)$ replaced by $\pi(\mu,\, \phi)$. Note that, in our case, we consider the posterior predictive $p(\boldsymbol{y}^*|\boldsymbol{y})$ to be for n random variables $Y_1^*, \ldots, Y_n^*$ having the same values of $m_1, \ldots, m_n$ as in the original variables $Y_1, \ldots, Y_n$.

To simulate values from $p(\boldsymbol{y}^*|\boldsymbol{y})$ we make use of the model structure involving $f(\boldsymbol{y}|\theta)$ and $g(\theta|\mu,\, \phi)$. The Metropolis algorithm described previously produces values of $(\mu,\, \phi)$ from the posterior $p(\mu,\, \phi|\boldsymbol{y})$. Because

$$h(\boldsymbol{y}|\mu,\, \phi) = \int f(\boldsymbol{y}|\theta)\, g(\theta|\mu,\, \phi)\, d\theta,$$

we may easily simulate from the posterior predictive using the following algorithm.

1. For each value of $(\mu,\, \phi)$ simulated from the posterior of these quantities, transform to $(\alpha,\, \beta)$ using the relation in expression (5).

2. Simulate a set of values $\theta_1, \ldots, \theta_n$ from $g(\theta|\alpha, \beta)$ which can be accomplished with the built-in R function for generating observations from a beta distribution.

3. For each set of $\theta_1, \ldots, \theta_n$ resulting from step 2, simulate a set of values $y_1^*, \ldots, y_n^*$ from binomial distributions with parameters θ_i and binomial "sample sizes" m_i; $i = 1, \ldots, n$.

4. Repeating this procedure for each value of (μ, ϕ) simulated from the posterior, we end up with M sets of data $\{y_i^* : i = 1, \ldots, n\}_k$; $k = 1, \ldots, M$ simulated from the posterior predictive distribution.

Model Assessment Using Posterior Predictive Values

Having simulated data sets from the posterior predictive distribution, it remains to use these values in a model assessment. A simple and effective strategy recommended by Gelman, Carlin, Stern and Rubin (1995) is based on the idea we have seen before that an adequate model should generate data similar to those we actually have. To put this idea into practice here, we choose one or more "features" of the data that indicate something meaningful about the way the model represents the situation. For example, we may look at the number of "extreme" values in data sets, the range or interquartile range, or perhaps the realized mean-variance relation. One caution is that functions of the data that correspond to sufficient statistics in the data model are probably not too valuable here, since these are how the data model influences the prior and, hence, are likely to be well represented by the posterior in all but the most hopelessly inadequate models.

Given a feature of the data of interest, represented as a function $q(y_1, \ldots, y_n)$ say, we may compute this function for each of the simulated data sets and compare the location of the value obtained from the actual data within the empirical distribution of these values. Note that this is similar to what we have already done with non-Bayesian analyses. There is, however, one important difference. If we simulate from a fitted model using maximum likelihood estimates (for example) as parameter values, we have simulated data from only one point in the parameter space. In the Bayesian approach using posterior predictive distributions we have simulated from a model that integrates over the entire parameter space.

Simulated Example

To illustrate this overall procedure, a simulated data set with $n = 25$, all $m_i = m = 20$, and parameters $\alpha = 2$, $\beta = 4$ (or $\mu = 0.33$, $\phi = 0.143$) was generated from a beta-binomial mixture model. The algorithm outlined above was run using a burn-in of $B = 50$ and a total of $M = 50,000$ kept values. Starting values were $\mu_0 = 0.5$ and $\phi_0 = 0.5$.

A histogram of $50,000$ simulated values from the posterior of μ is presented in Figure 1, and a histogram of the same number of simulated values from the posterior of ϕ is presented in Figure 2. True parameter values are shown with solid vertical lines (although they can be easy to miss in these figures). Posterior expectations were 0.293 for μ and 0.0368 for ϕ.

An assessment of model adequacy was conducted by simulating $50,000$ data sets from the posterior predictive distribution as described previously. Three quantities were chosen to reflect features of these data sets, the mean observed proportion, the variance of observed proportions, and the difference between the maximum and minimum proportions (i.e., the range of observed proportions). Histograms of values for these three quantities obtained from the simulated data sets are presented in

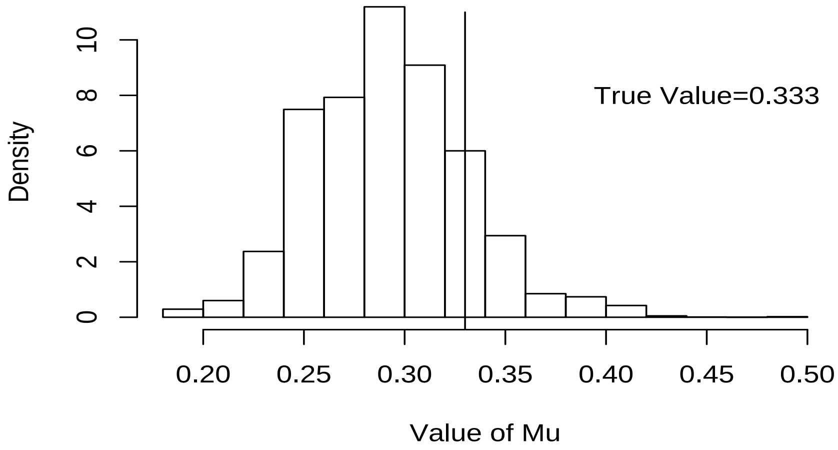

Figure 1: Posterior Distribution of μ From Simulated Example

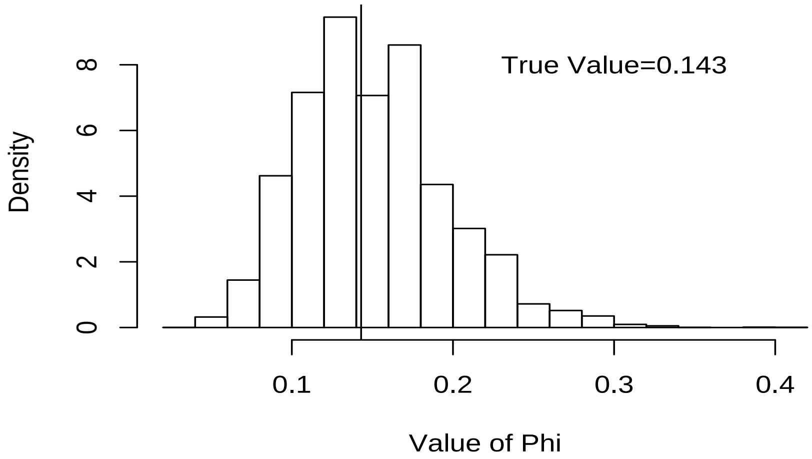

Figure 2: Posterior Distribution of ϕ From Simulated Example

Figures 3, 4, and 5, along with the values corresponding to the actual observed data (which in this case was also simulated).

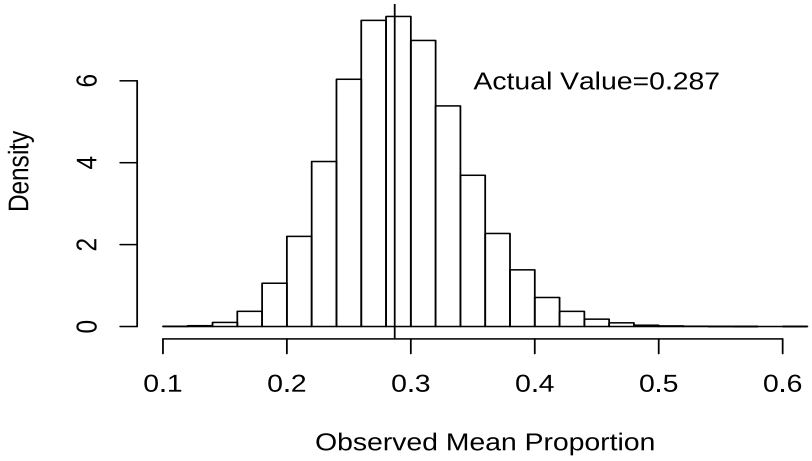

Figure 3: Distribution of mean proportions from simulated data sets

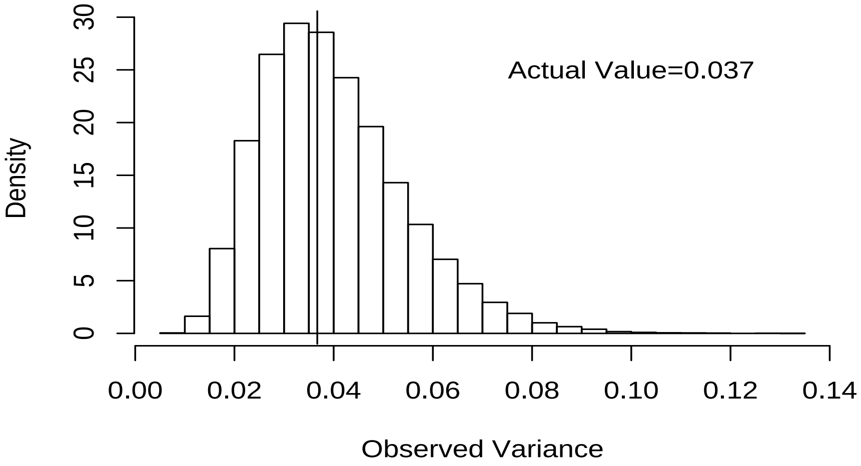

Figure 4: Distribution of variances of proportions from simulated data sets

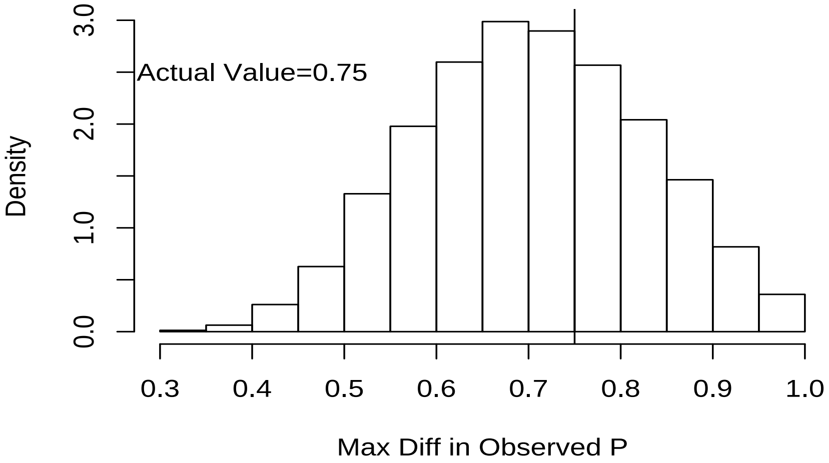

Figure 5: Distribution of range in proportions from simulated data sets

Gambusia Reproduction Study

The same model developed above was applied to the study of reproductive success (or failure) in *Gambusia* in the Central Valley of California. This study is described in Example 7.12 in the course notes, and an analysis based on maximum likelihood is presented in Example 8.13. Here, we use the Bayesian analysis for a beta-binomial mixture model with data from the San Luis drain for which fish were held in the same water from which they were collected. The raw data are given in the table that accompanies Example 8.13.

The same Metropolis algorithms as used in the simulated example was applied to these data, with $B = 50$, $M = 50,000$, and initial values $\mu_0 = 0.5$, $\phi_0 = 0.5$. Histograms of the posterior distribution of μ is presented in Figure 6 and that for ϕ is presented in Figure 7. Table 1 compares maximum likelihood estimates for these data with the corresponding values from the Bayesian Analysis.

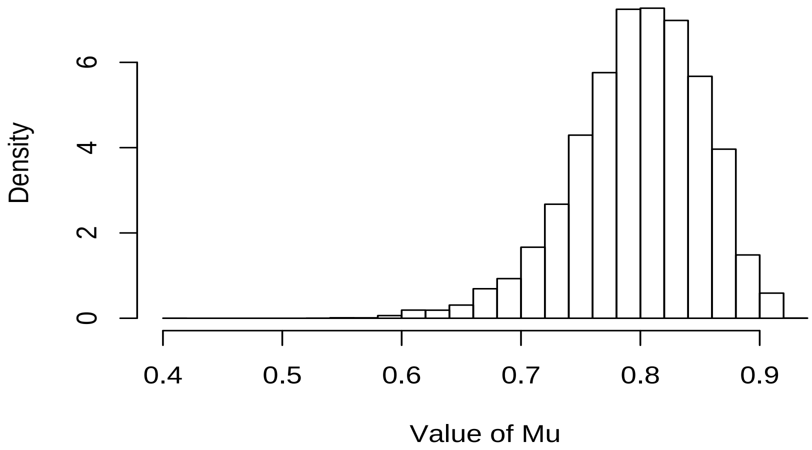

Figure 6: Posterior Distribution of μ From *Gambusia* Reproduction Study

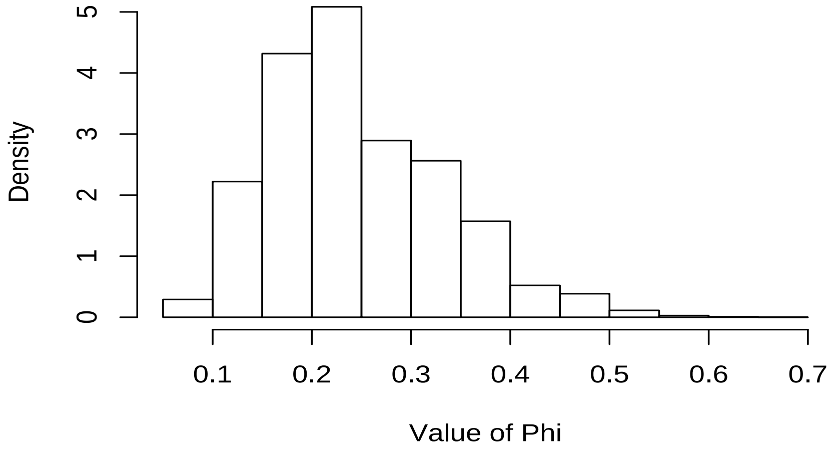

Figure 7: Posterior Distribution of ϕ From *Gambusia* Reproduction Study

The values in Table 1 for maximum likelihood analysis of ϕ are not the same as those presented in the table of Example 8.13. In that analysis, we took $\mu = \alpha/(\alpha+\beta)$ the same as here, but we defined $\phi = 1/(\alpha + \beta)$. Here, we used $\phi = 1/(\alpha + \beta + 1)$. The maximum likelihood point estimate of $\hat{\phi} = 0.244$ (see Example 8.13 in the course notes) was transformed to be comparable with the definition of ϕ used here in the Bayesian analysis. Specifically, if $\hat{\phi}$ is the maximum likelihood estimate of

	Maximum Likelihood			Bayes		
Parameter	Estimate	Variance	90% Interval	Estimate	Variance	90% Interval
μ	0.822	0.0023	(0.744, 0.901)	0.799	0.0029	(0.702, 0.878)
ϕ	0.196	0.0091	(0.087, 0.401)	0.245	0.0078	(0.123, 0.403)

Table 1: Comparison of maximum likelihood and Bayes estimates for the *Gambusia* reproduction study.

$1/(\alpha + \beta)$ then $\phi/(\phi + 1)$ is the maximum likelihood estimate of $1/(\alpha + \beta + 1)$ and this is the value in Table 1 (0.196). The value for variance and the interval estimate in Table 1 were produced using the delta method since in the maximum likelihood analysis the interval was produced using Wald theory. The values in Table 1 show overall agreement between the maximum likelihood and Bayesian analyses, which we would expect. The Bayesian interval for μ is shifted slightly to the left (smaller) from the maximum likelihood interval, and is also a bit wider (0.176 for the Bayesian interval versus 0.157 for the mle interval). The Bayesian point estimate of ϕ is greater than the mle, but there is less difference in the intervals that this might suggest. Interestingly, the Bayesian interval is more narrow (width of 0.280) than is the Wald theory interval (width of 0.314).

The same model assessment used with the simulated example was conducted for these data, resulting in Figure 8 for means of observed proportions, Figure 9 for variances of observed proportions, and Figure 10 for range of observed proportions.

If one computes posterior predictive p-values from the values shown in these figures, one obtains $p = 0.3800$ for means (i.e., the proportion of means from simulated data sets that are greater than the value from the actual data is 0.38). The p-value for variances is $p = 0.3791$, and that for ranges is $p = 0.2565$. In all, these values indicate little in the way of model deficiency. The spread of the ranges exhibited in Figure 10 is perhaps interesting, but careful thought would be needed before concluding too much from this phenomenon.

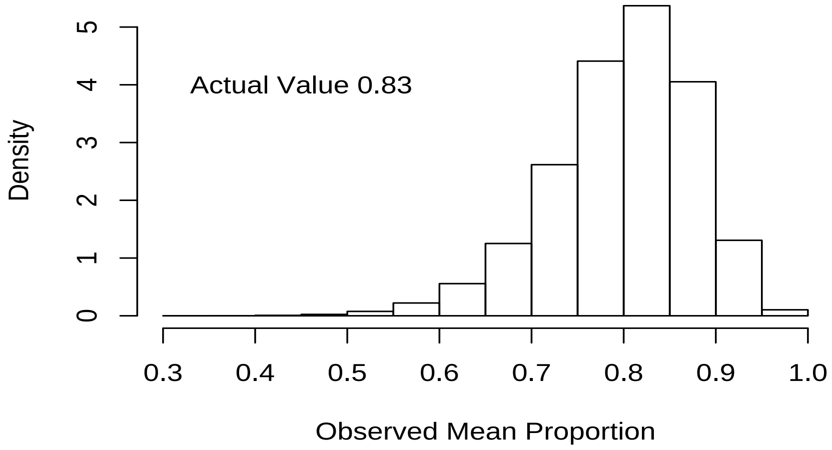

Figure 8: Distribution of mean proportions from simulated data sets for the *Gambusia* study

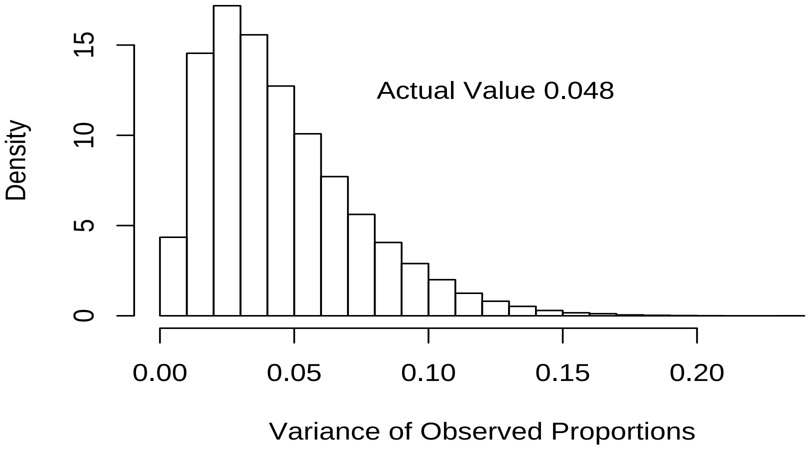

Figure 9: Distribution of variances of proportions from simulated data sets for the *Gambusia* study

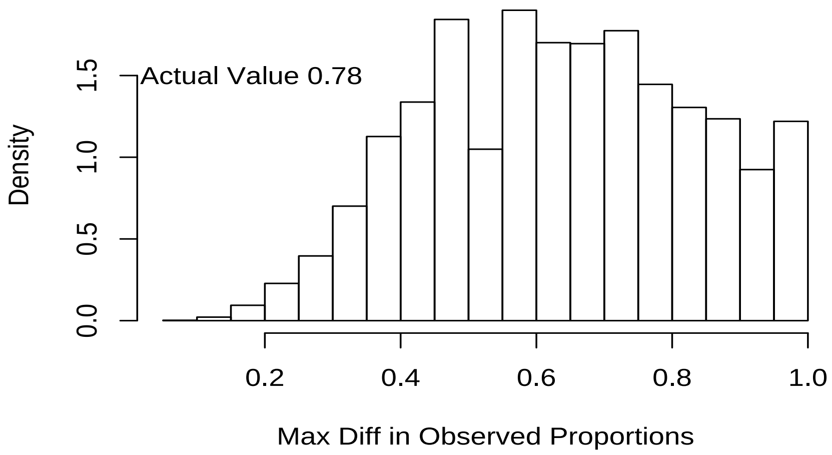

Figure 10: Distribution of range in proportions from simulated data sets for the *Gambusia* study