Model-based estimation of effective sample size in Stock Synthesis using the Dirichlet-Multinomial distribution

James T. Thorson1, Kelli F. Johnson2, Richard D. Methot1, Ian G. Taylor1

1Fisheries Resource Assessment and Monitoring Division, Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 2725 Montlake Blvd. East, Seattle, WA 98112, USA

2School of Aquatic and Fishery Sciences, University of Washington, Box 355020, Seattle, WA 98195-5020, USA

**Abstract**

Theoretical considerations and applied examples suggest that stock assessments are highly sensitive to the weighting of different data sources whenever data sources conflict regarding parameter estimates. Previous iterative reweighting approaches to weighting compositional data are generally ad hoc, do not propagate uncertainty about data-weighting when calculating uncertainty intervals, and often are not re-adjusted when conducting sensitivity or retrospective analyses. We therefore incorporate the Dirichlet-Multinomial (DM) distribution into Stock Synthesis, and propose it as a model-based method for estimating effective sample size. This distribution incorporates one additional parameter per survey (with the option of mirroring its value among fleets), and we show that this parameter represents the ratio of nominal (“input”) and effective (“output”) sample size. We demonstrate this approach using data for Pacific hake, where DM and McAllister-Ianelli reweighting approaches give similar and plausible results. We also use simulation testing to demonstrate the estimation properties of this new estimator, and show that it provides approximately unbiased estimates of variance inflation when compositional samples capture clusters of individuals with similar ages/lengths. We conclude by recommending further research to develop computationally efficient estimators of effective sample size that are based on alternative, *a priori* consideration of sampling theory.

**Keywords**: Dirichlet-Multinomial; integrated assessment; multinomial; statistical catch-at-age

**Introduction**

Stock assessment models are quantitative tools that are used to provide a scientific basis for the management of marine fishes (Walters and Martell, 2004). Assessment models increasingly incorporate biological assumptions regarding the population dynamics of fished species, and population dynamics parameters are estimated by fitting the assessment model to available data (Maunder and Punt, 2013). Fitting population models to available data is typically done using likelihood-based statistics, and the proper estimation of confidence and forecast intervals therefore generally requires accounting for correlated residuals as caused by unmodeled biological or measurement process (Thorson and Minto, 2015). Theoretical considerations and applied examples suggest that integrated statistical stock assessments are highly sensitive to the weighting of different data sources whenever sources conflict regarding parameter estimates. Consequently, estimates of stock status and productivity are often highly dependent upon the weighting of different data sources (Francis, 2011).

Stock assessment models frequently include sampling data regarding the proportion of the vulnerable population that belong to different observable categories. Common types of categorical information include the proportion of different ages, lengths, and sexes. In practice, this compositional information arises from a process of sampling fishes (e.g., non-extractive visual samples, or by capturing and measuring fishes), and samples must then be standardized (sometimes termed “expanded”) to account for distribution of sampling effort to collect compositional data (Thorson, 2014). Compositional standardization ideally results in an estimate of variance (e.g. effective sample size) for the compositional data in a given year that then is used in the stock assessment model. However, unaccounted aspects of the sampling contribute additional uncertainties (Crone and Sampson, 1998) such that the direct estimate of sample size is best considered as the “input” sample size that should be adjusted to attain the actual “effective” sample size.

For compositional data, which describe the distribution of ages or lengths in catches, the multinomial distribution is the most used. The multinomial distribution assumes samples are collected at random from the population. Therefore, employed sampling methods and fish behavior (e.g., schooling) will lead to overdispersion of the true uncertainty in the estimated proportions (citation).

* Assumptions and violations.
* Use of the multinomial requires determining an effective sample size.
  + Methods for data weighting should allow for correlations (Francis, 2011).

Alternatives to the multinomial, including ad hoc practices.

* McAllister-Ianelli (1997)
* Bootstrap (Stewart and Hamel, 2014)
* Dirichlet-Multinomial (DM)
* Logistic-normal (Francis, 2014)
* Adjusted log-normal (Legault, 2014)
* Multivariate-logistic (Schnute and Richards, 1995)

Given the inability of *ad hoc* methods to propagate uncertainty about data-weighting when calculating uncertainty intervals and their propensity to encourage ignoring data-weighting when conducting sensitivity or retrospective analyses, we incorporate the DM distribution into Stock Synthesis (SS) and propose it as a model-based method for estimating the effective sample size ().

**Methods**

*Introducing the Dirichlet-multinomial distribution*

We here use a Dirichlet-Multinomial distribution:

where is the proportion at age in the available data such that , *N* is the total number of samples in the available data (which is restricted to any non-negative real number), is the estimated proportion at age (such that ), and is the estimated variance inflation coefficient (we use the gamma function, rather than the conventional factorial function, so that the Dirichlet-Multinomial is defined for all non-negative sample sizes *N*). The first term does not depend upon the parameters, but ensures that as , :

i.e., that the Dirichlet-multinomial reduces to the multinomial likelihood as the variance-inflation coefficient goes to infinity.

*Computing the effective sample size*:

We define the effective sample size of a distribution for compositional data *c* as the sample size of a multinomial distribution that has the same variance. The variance of a single element from a multinomial distribution is

where *N* is the sample size. Defining observed proportion , we see that:

i.e., variance decreases as the reciprocal of sample size.

We next introduce a Dirichlet distribution:

where and is the true proportion at age. The Dirichlet distribution has variance:

such that is the effective sample size of the Dirichlet distribution:

Finally, we the variance of the observed proportion at age for a Dirichlet-multinomial distribution is approximately:

We therefore calculate the estimated effective sample size *Neff* of a Dirichlet-multinomial distribution as:

i.e., that the effective sample size is the harmonic sum of *N* and .

*Two potential parameterizations*

Given the Dirichlet-mulinomial distribution and the closed-form computation of its effective sample size, we propose two alternative parameterizations that may be useful in practice for length and age-composition samples in stock assessment models. These parameterizations differ in terms of the function relating input and effective sample size (Fig. 1).

*Parameterization #1 – Linear effective sample size*

As a default, we recommend a re-parameterizations of the Dirichlet-multinomial distribution, wherein the variance-inflation parameter is replaced by a linear function of input sample size *N,* i.e., . This results in the following probability distribution function:

which has effective sample size:

Given that *N>>1* and <<*1*, this reduces to:

i.e., the parameter is the ratio of effective and input sample size. We recommend using the “linear effective sample size” parameterization given that previous methods for weighting compositional data have generally multiplied the likelihood of compositional data by a fixed quantity *λ<1*, and this parameterization has similar behavior for large input sample sizes (in practice *N>10*).

*Parameterization #2 – Asymptotic effective sample size*

As a potential alternative, analysts may instead use the original parameterization of the Dirichlet-multinomial distribution:

with effective sample size:

This parameterization can revert to the multinomial distribution with sufficiently large , i.e., when .However, it provides an upper bound on effective sample size with lower values of , i.e., when . Therefore, this parameterization could be useful whenever analysts seek to estimate an upper bound on the effective sample size for a given year.

We have implemented both parameterizations of the DM distribution in SS (version 3.30; release date Jan, 2016), an integrated age-structured stock assessment framework frequently used to conduct assessments in many parts of the world (Methot and Wetzel, 2013). In the following, however, we focus exclusively on the linear parameterization (version #1), given that it more closely mimics the action of the McAllister-Ianelli iterative reweighting method. We leave for future research the development and exploration of more-complicated two-parameter forms for the Dirichlet-multinomial distribution, which could combine the characteristics of both versions.

*Case study: Pacfic hake*

To demonstrate this new data-weighting method, we compare its performance with a recent stock assessment for Pacific hake (*Merluccius productus*). Pacific hake is a semi-pelagic schooling species of commercial importance to fisheries off of the US West Coast and Western Canada. Recent management includes an international treaty informed by annual stock assessments conducted using SS. Data used in the assessment includes catches from 1966 to 2014, fishery age composition, ten intermittent acoustic surveys conducted between 1995 and 2013, with associated survey age-composition samples, and ‘empirical’ fishery weight-at-age data, which are assumed to be known without error (Taylor et al., 2015).

Four assessment models were fit to data for Pacific hake, where each model used a different approach to data-weighting for the fishery: (i) unweighted, (ii) tuned using McAllister-Ianelli (1997), (iii) estimated using the Dirichlet-multinomial distribution, and (iv) weight of zero for the fishery age-composition data. Options (ii) (“McAllister-Ianelli”) involved fitting the model to simulated data, computing the ratio of input and “effective” sample size for each year of age-composition data for the fishery, computing the harmonic mean ratio among years, and then inputing this value as the a “weighting factor” for the fishery-age composition data. The model was then fitted to data a 2nd time, this process was repeated, and the third fit to data was used as the final estimate of parameters. Option (iv) specifies that the stock assessment is fitted only to abundance indices and survey age-composition data, and represents the extreme case of “zero” weight assigned to fishery compositional data. Preliminary exploration showed that the input sample size is approximately equal to effective sample size for survey age-composition data (i.e., the McAllister-Ianelli approach results in a ratio of 0.94, and the Dirichlet-multinomial results in a ratio of 1.00. We therefore chose to not down-weight the survey age-composition data (i.e., we did not estimate the Dirichlet-multinomial parameter for the survey, nor did we tune it).

*Simulation testing*

The performance of the Dirichlet-multinomial distribution implemented in Stock Synthesis was explored using simulated data. To do so, we simplified the Pacific hake estimation model in four ways: (1) we changed fishery selectivity to be stationary over time (i.e., removing time-varying selectivity parameters), (2) we changed all fishery age-composition sample sizes to a single fixed value in each year, (3) we changed all survey age-composition sample sizes to 100 samples per year, and (4) we changed to using an “explicit-F” parameterization, wherein instantaneous, fully-selected fishing mortality in each year is estimated as a fixed effect. We then ran the modified Pacific hake assessment model on available data, extracted estimated parameters, and used these estimates as the “true” value during the simulation experiment (while confirming that estimated stock status and productivity is generally similar to that in the case study).

We then generate new, simulated data sets using the Stock Synthesis parametric bootstrap simulator. For each simulation replicate, we simulate a new vector of recruitment deviations, and also simulate a new fishing mortality pattern that increases linearly from *F=0.01* in the first year (1966) to *F=0.30* in the final year (2013). The bootstrap simulator then calculates the population abundance-at-age resulting from the input vector of recruitment deviations and fishing mortality, and simulates abundance index and age-composition samples from their specified distributions (i.e., using a lognormal distribution with the input log-standard deviation for the abundance index, and a multinomial distribution with the input sample size for the age-composition samples).

The simulation experiment involves a factorial design with three simulation scenarios, four levels of an inflation factor, and three estimation models. For each combination, we run 100 simulation replicates, for a total of total estimation model runs. We define three simulation scenarios, where the “true” sample size used to simulate age-composition data for the fishery is either 25, 100, or 400 per year. Given this “true” age-composition sample, the “observed” age-composition sample is then inflated by a fixed inflation factor *θsim*, with value . We then use estimation models (i), (ii), and (iii) defined in the section titled *Case study: Pacfic hake* (see above).

*Model evaluation*

Estimation procedures were evaluated by comparing estimated parameters and derived quantities of interest to management to their true values from as defined in the operating model. Estimation error was quantified using relative error (, where and are estimated and true parameter values respectively).

Explain ESS across models compared to the true.

**Results**

*Case study application: Pacific hake*

Comparing four alternative methods for weighting compositional data in the Pacific hake assessment (Fig. 2) shows that estimates of spawning output and fishing intensity are generally bracketed by the two naïve approaches, i.e., either treating input sample size as effective sample size (“no weighting”) or removing fishery age-composition data entirely (“no fishery ages”). In particular, removing fishery age data results in a higher estimate of average unfished spawning output and lower spawning output estimates from the mid-1980s onward, while treating input as effective sample size results in strong year-class strength estimates in the early 1980s and early 2000s. By contrast, the default McAllister-Ianelli and new Dirichlet-multinomial weighting methods results in similar estimates of spawning output, with the exception of recent years (2010 onwards) when the Dirichlet-multinomial estimator results in somewhat elevated estimates of spawning output relative to the McAllister-Ianelli method. Similarly, the McAllister-Ianelli and Dirichlet-multinomial estimates of fishing intensity are more similar than the other weighting methods, particularly for early years (prior to 1970).

*Simulation experiment*

Estimates of the Dirichlet-multinomial parameter are hugely different among the three scenarios and three levels of the inflation factor (Fig. 3, panel a). However, estimates of effective sample size are generally similar for all levels of the inflation factor for a given scenario (Fig. 3, panel b). In general, the estimated effective sample size closely matches the true sample size for all scenarios and levels of the inflation factor. However, we detect a small positive bias in the estimates of effective sample size when the true sample size is 400 (i.e., median effective sample size estimate is close to 450). This bias is not detected in the other scenarios (true sample size of 25 or 100).

Comparison of parameter estimates from the unweighted multinomial, McAllister-Ianelli reweighting algorithm, and the linear parameterization of the Dirichlet-multinomial distribution shows that the McAllister-Ianelli and Dirichlet-multinomial have similarly precision and accuracy when estimating natural mortality and average unfished recruitment for all levels of the inflation factor. By contrast, the unweighted model has seriously substantially degraded estimates of natural mortality and unfished recruitment for any inflation factor other than 1. We note that the Dirichlet-multinomial algorithm has a notable fraction (20 of 100) of replicates that do not converge for the highest level of the variance inflation (*θsim*=1000), but convergence is otherwise comparable between methods. We therefore conclude that the Dirichlet-mulinomial method has similar estimation performance to the previous McAllister-Ianelli reweighting approach.

**Discussion**

In this paper, we have shown that the Dirichlet-multinomial distribution can be used to generate model-based estimates of effective sample size for age and length-compositional data in stock assessment models. For this purpose, we have implemented two parameterizations of the Dirichlet-multinomial distribution in the widely-used Stock Synthesis software. We have then applied the model to data for Pacific hake, showing that it provides estimates in agreement with the previous McAllister-Ianelli approach, and provide a simulation experiment to verify that it provides unbiased estimates of effective sample size given that the model is otherwise specified correctly.

We believe that the Dirichlet-multinomial approach is superior to alternative data-weighting methods for several reasons.

1. *Slow or inconsistent exploration of alternative models*: Previous methods (e.g., the McAllister-Ianelli method) require fitting a stock assessment model to data, extracting residuals, estimating effective sample size estimates from this fit, and then re-estimating the model. This iterative tuning procedure either slows exploration of alternative models (due to the need for re-tuning after each model change) or causes inconsistent exploration of alternative models (where analysts neglect to re-tune for every sensitivity run, and therefore compare between runs that are not tuned in a consistent manner).
2. *Failure to account for uncertainty in data weighting*: Previous methods also provide no obvious method for propagating uncertainty about data-weighting. By contrast, the Dirichlet-multinomial approach represents data-weighting via an estimated parameter, and the uncertainty in this parameter can be captured via standard statistical methods (e.g., likelihood profiles, asymptotic confidence intervals, or Bayesian posteriors, (Magnusson et al., 2013)).
3. *Clear standards for convergence*: Previous methods also require a subjective decision regarding when to stop tuning the sample size, what order to tune multiple fleets, and how to combine data-weighting information from multiple fleets. This subjective decision is rarely documented, and different decisions by different analysts may cause substantial differences in ultimate estimates of stock status and productivity in assessments where data weighting is an important axis of uncertainty (e.g., sablefish). By contrast, the Dirichlet-multinomial method allows for a single, unambiguous definition of convergence (i.e., via maximizing the model likelihood), which can be independently replicated by different authors and does not require further documentation. Given our observation of a 20% non-convergence of the D-M method when the sample size adjustment ratio needed to be extreme (1000), a possible approach would be to do one model run using the M-I approach to get a good initial value for the D-M adjustment parameter, then proceeding to fully estimate that parameter in a final model run.
4. *Interpretable estimates of effective sample size*: Analysts have previously suggested altnerative model-based methods for estimating effective sample size. For example, an analyst might use a simple Dirichlet distribution, rather than the Dirichlet-multinomial distribution used here. However, the Dirichlet distribution can have effective sample size that ranges from 0 to infinity, i.e., it can exceed the input sample size. By contrast, the Dirichlet-multinomial distribution ensures that the effective sample size can never be greater than the input sample size.

In particular, we envision that benefit #4 (“interpretable estimates of effective sample size”) can be used as a diagnostic for model goodness-of-fit. Specifically, we envision that the analyst can subsequently explore potential hypotheses for overdispersed compositional data when the effective sample size is lower than the input sample size. Potential causes presumably include time-varying or non-parametric fishery selectivity, time-varying growth, and other common types of model misspecification. The analyst could then sequentially additional flexibility in these processes by treating them as random effects (Thorson et al., 2015), and could determine which change causes the magnitude of overdispersion to decrease.

**Acknowledgements**

Partial support for KFJ was provided by the National Marine Fisheries-Sea Grant Population Dynamics Fellowship (?). This publication was partially funded by the Joint Institute for the Study of the Atmosphere and Ocean (JISAO) under NOAA Cooperative Agreement No. ?, Contribution ?.

**References**

Francis, R.I.C.C., 2011. Data weighting in statistical fisheries stock assessment models. Can. J. Fish. Aquat. Sci. 68, 1124–1138.

Magnusson, A., Punt, A.E., Hilborn, R., 2013. Measuring uncertainty in fisheries stock assessment: the delta method, bootstrap, and MCMC. Fish Fish. 14, 325–342.

Maunder, M.N., Punt, A.E., 2013. A review of integrated analysis in fisheries stock assessment. Fish. Res. 142, 61–74.

Thorson, J.T., 2014. Standardizing compositional data for stock assessment. ICES J. Mar. Sci. J. Cons. 71, 1117–1128. doi:10.1093/icesjms/fst224

Thorson, J.T., Hicks, A.C., Methot, R.D., 2015. Random effect estimation of time-varying factors in Stock Synthesis. ICES J. Mar. Sci. J. Cons. 72, 178–185. doi:10.1093/icesjms/fst211

Thorson, J.T., Minto, C., 2015. Mixed effects: a unifying framework for statistical modelling in fisheries biology. ICES J. Mar. Sci. J. Cons. 72, 1245–1256. doi:10.1093/icesjms/fsu213

Walters, C.J., Martell, S.J.D., 2004. Fisheries Ecology and Management. Princeton University Press, Princeton, New Jersey.

Francis, R.I.C.C. 2011. Data weighting in statistical fisheries stock assessment models. Can. J. Fish. Aquat. Sci. 68, 1124-1138.

Francis, R.I.C.C. 2014. Replacing the multinomial in stock assessment models: a first step. Fish. Res. 151, 70-84.

Hulson, P.-J.F., Hanselman, D.H., Quinn, T.J., II. 2012. Determining effective sample size in integrated age-structured assessment models. ICES J. Mar. Sci. 69, 281-292.

Legault, C.M. 2014. The ability of two age composition error distributions to estimate selectivity and spawning stock biomass in simulated stock assessments. Fish. Res. 158, 172-180.

McAllister, M.K., Ianelli, J.N. 1997. Bayesian stock assessment using catch-age data nd the sampling – importance resampling algorithm. Can. J. Fish. Aquat. Sci. 54, 284-300.

Methot, R.D., Wetzel, C.R. 2013. Stock Synthesis: a biological and statistical framework for fish stock assessment and fishery management. Fish. Res. 142, 86-99.

Schnute, J. T., Richards, L. R. 1995. The influence of error on population estimates from catch-age models. Can. J. Fish. Aquat. Sci. 52, 2063-2077.

Stewart, I.J., Hamel, O.S. 2014. Bootstrapping of sample sizes for length- or age-composition data used in stock assessments. Can. J. Fish. Aquat. Sci. 71, 581-588.

Taylor, I.G., Grandin, C., Hicks, A.C., Taylor, N., Cox, S. 2015. Status of the Pacific Hake (whiting) stock in U.S. and Canadian waters in 2015. Prepared by the Joint Technical Committee of the U.S. and Canada Packfic Hake/Whiting Agreement; National Marine Fishery Service; Canada Department of Fisheries and Oceans. 159 pp.

Table 1. Life-history, fishery, and modelling parameters used for the simulation.

Table 2. Combinations of operating and estimating models executed for the simulation.

Fig. 1. Input sample size (x-axis) and effective sample size (; y-axis) for two paramaterizations of the Dirichlet-Multinomial (DM) distribution across varying values for the DM parameter specific to each parameterization. The dashed line represents the 1:1 line where the input sample size is the same as the .

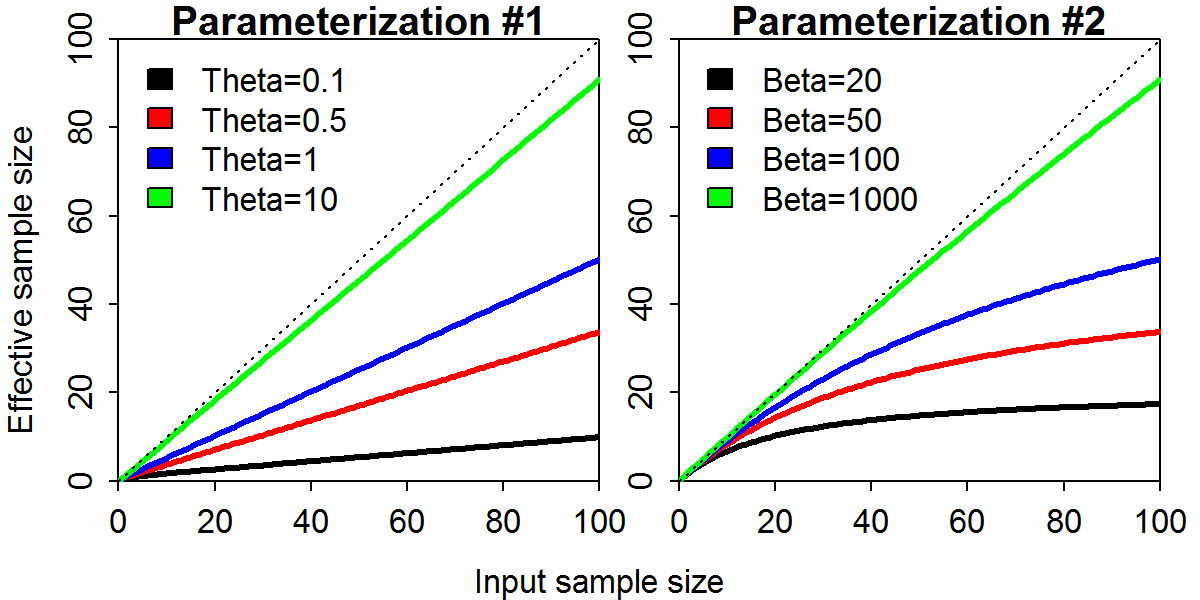
****

Fig. 2. Comparison of spawning output relative to average unfished levels (top), spawning output (SPB; middle), and fishing intensity (FSPR; right) for the Pacific hake assessment given four alternative methods of weighting the age-composition data: (i) unweighted (red), (ii) tuned (green); (iii) Dirichlet-Multinomial distribution (blue); and (iv) weight of zero for the age-composition data (black), where for each modle we show the maximum likelihood estimates (solid line) and +/- 1 standard error (shaded region).

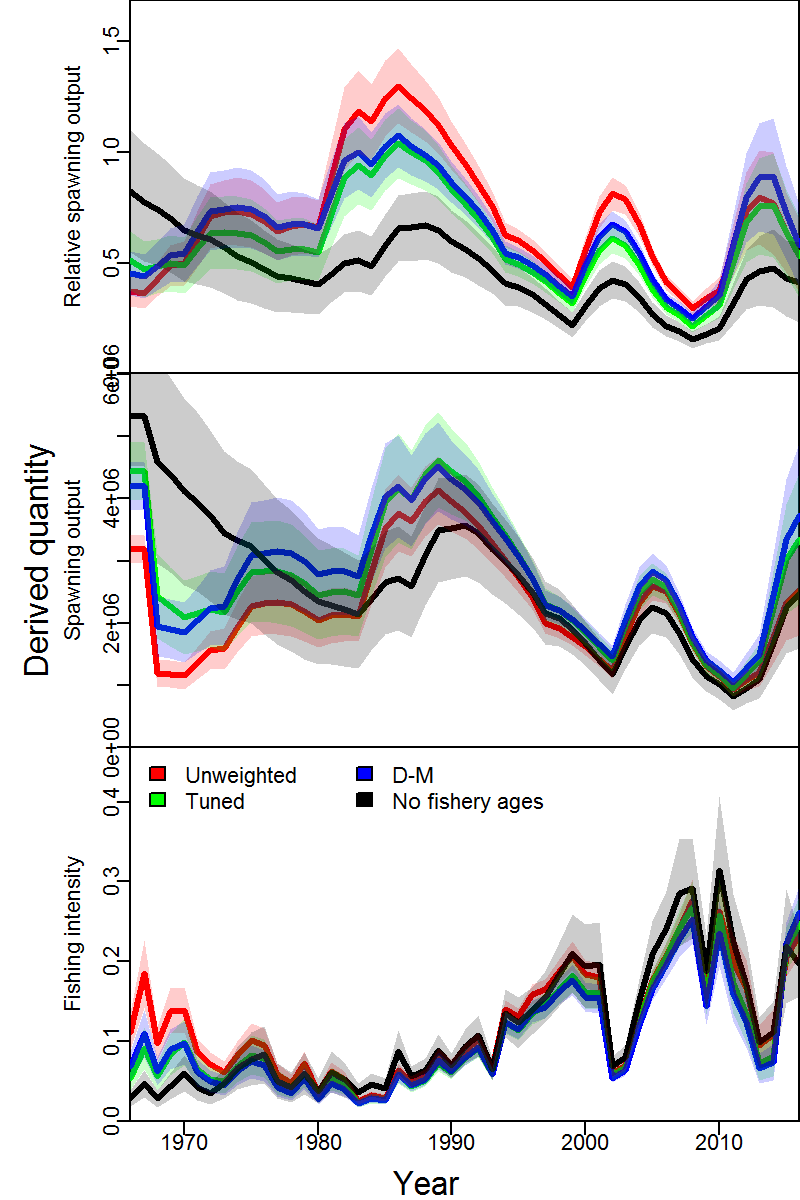


Fig. 3. Estimated Dirichlet-multinomial variance inflation parameter (top panel) and effective sample size (, bottom panel) from the “linear” parameterization (version #1) of the Dirichlet-Multinomial (DM) distribution implemented in Stock Synthesis shown for three “true sample sizes” (representing an increase in available information from 25 to 400 true samples per year) and three levels of variance inflation (wherein the input sample size provided to Stock Synthsis is 25, 100, or 1000 the true sample size).

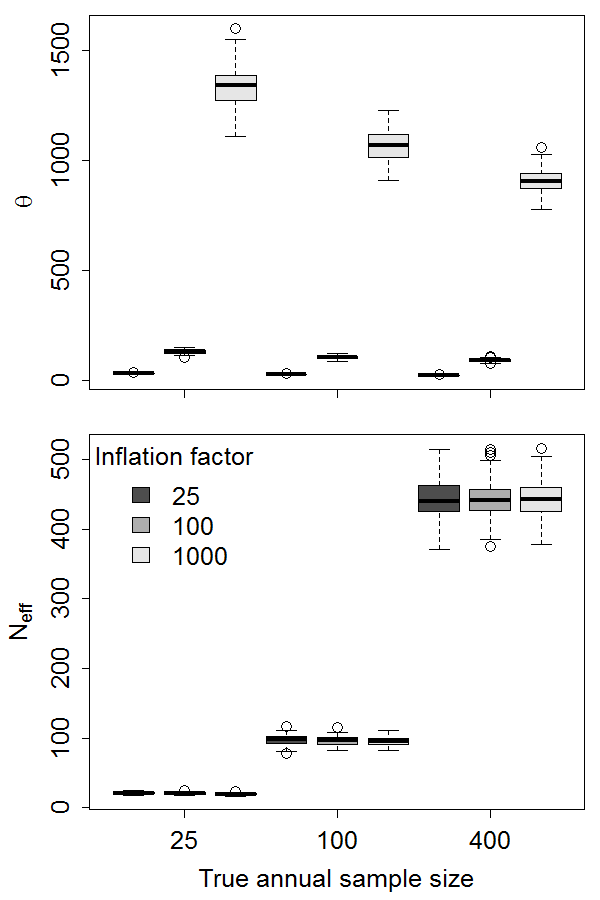


Fig. 4. Relative error in parameter estimates across estimation methods (rows; “M-I”: reweighted using the McAllister-Ianelli method; “unweighted”: conventional multinomial treating input as effective sample size; “D-M”: linear-parameterization of the Dirichlet-multinomial distribution) and levels of the inflation factor for the fishery age-composition data in the operating model (columns). Each panel depicts the maximum likelihood estimates of natural mortality rate (*M,* y-axis) and average unfished recruitment (*ln(R0)*, x-axis), where colors are used to distinguish estimates. We only show results for estimation models where the maximum final gradient was <0.1 (the number of replicates across models is indicated in each panel, where 300 implies that all 100 replicates coverged for each of three estimation models), and confirm that results are qualitatively similar if using a different convergence threshold. The lower left panel is not plotted because the DM estimation method was not used when the inflation factor was one.

