Technical Specification for function N_at_Age.
Supplied by Marcel Machiels. 4 December 2008.

Type 3 sampling

Type 3 sampling allows assessment of length per fish sampled for each market category. Numbers-at-length are estimated via multiplication with a raising factor, which is defined as the inverse of the biomass fraction sampled. This is done in two steps, i.e. a double raising procedure is performed. First, observed numbers are raised to a total per category per sampled landing, taking into account the differences in biomass fraction sampled per category per landing. Next, numbers-at-length are raised to the total landings. This involves multiplication with the ratio of total landing weight to total weight of sampled landings, either by market category [Type 3a] or for all categories combined [Type 3b].

Numbers-at-age are estimated via conversion of the length distribution into an age-distribution according to an age-length key. An age-length key is constructed on the basis of a subset, which is stratified by length classes, ideally such that every length class is represented by equal numbers. Possible differences between market categories or sampled landings are not considered.

Mathematical notation

Let $c \in \{1, 2, ..., C\}$ be the indicator for market category and let $s \in \{1, 2, ..., S_c\}$ be the indicator for a sampled landing. Because a sampled landing needs not necessarily contain all market categories, the total number of samples may vary between categories. In Type 3a sampling, the total number of fish is the summed product of F_c, the per category ratio of total landing weight to the total weight of sampled landings, and n_c, the number of fish per category summed over all sampled landings. Else, it is the product of F_o, the overall ratio of total landing weight to the total weight of sampled landings, and $\Sigma_c\, n_c$, the number of fish summed over all categories and all sampled landings.

The aggregate landing weight per market category W_c is considered a known figure in Type 3a sampling. Let $w_{c,s}$ denote the weights per category of a sampled landing and $\overline{w}_c$ the average landing weight of a category, then F_c can also be expressed as

[1a] $$F_c = \frac{W_c}{\Sigma_s w_{c,s}} = \frac{W_c}{S_c \overline{w}_c}$$

In Type 3b sampling, only the total landing weight W is known and the raising factor is

[1b] $$F_o = \frac{W}{\Sigma_c \Sigma_s w_{c,s}} = \frac{W}{\Sigma_c S_c \overline{w}_c}$$

The number of fish per category per sampled landing $n_{c,s}$ is either observed directly, or estimated by multiplication of the number of fish sampled $n(s)_{c,s}$ with the ratio of $w_{c,s}$ to the total sample weight per category per landing sampled $w(s)_{c,s}$. In a sense, this comes down to using the proportion biomass sampled as a proxy for the probability of fish being sampled. If $P(s)_{c,s}$ denotes the sample proportion per category per sampled landing, then $n_{c,s}$ equals

[2] $$n_{c,s} = \frac{n(s)_{c,s} w_{c,s}}{w(s)_{c,s}} = \frac{n(s)_{c,s}}{P(s)_{c,s}}$$

From this, the number of fish per category summed over all sampled landings is estimated as

[3] $$n_c = \sum_s \frac{n(s)_{c,s}}{P(s)_{c,s}}$$

To partition the number of fish per category between various age classes $a \in \{1, 2, ..., A\}$, we should obtain an estimate of $P_{a,c}$, the probability that a fish is of a particular age given the market category it belongs to. Estimation of $P_{a,c}$ involves integration of the product of the age-length key $P_{a,l}$ and the length probability distribution per market category $P_{l,c}$ over the entire length range $l \in \{1, 2, ..., L\}$, or

[4] $$P_{a,c} = \sum_l P_{a,l} P_{l,c}$$

In Type 3a sampling, the numbers-at-age raised to the total landing N_a can then be estimated as

[5a] $$N_a = \sum_c F_c \, n_c \, P_{a,c} = \sum_c \frac{W_c \, n_c \, P_{a,c}}{S_c \, \overline{w}_c}$$

In Type 3b sampling, the total numbers-at-age are estimated as

[5b] $$N_a = F_o \sum_c n_c \, P_{a,c} = \frac{W \, \Sigma_c \, n_c \, P_{a,c}}{\Sigma_c \, S_c \, \overline{w}_c}$$

Statistical considerations

Let us assume that uncertainty is negligible in the aggregate landing weights. Equations [5a] and [5b] then illustrate that uncertainty in the total numbers-at-age is governed by inaccuracy in the estimation of (i) the average landing weight per market category, (ii) the number of fish per category summed over all sampled landings, and (iii) the probability that fish are of a particular age (given market category).

The uncertainty in the total numbers-at-age estimates can be approximated by application of the general formula for error propagation and is quantified as **dEREd′**. In this notation, **d** denotes a row vector of partial derivatives towards stochastic variables, **E** contains error terms in a diagonal structure and **R** is a symmetrical correlation matrix.

As each term in the summation of Equation [5a] contributes three stochastic variables, **d** has length $3C$. Applied to Equation [5a], the vector **d** looks like

$$[6] \qquad \left[\frac{-W_1\, n_1\, P_{a,1}}{S_1\, \overline{w}_1^{\,2}} \quad \frac{W_1\, P_{a,1}}{S_1\, \overline{w}_1} \quad \frac{W_1\, n_1}{S_1\, \overline{w}_1} \quad \cdots \quad \frac{-W_C\, n_C\, P_{a,C}}{S_C\, \overline{w}_C^{\,2}} \quad \frac{W_C\, P_{a,C}}{S_C\, \overline{w}_C} \quad \frac{W_C\, n_C}{S_C\, \overline{w}_C} \right]$$

The length of **d** determines the row and column dimensions of the matrices **E** and **R**. To fill **R** we must assume that the correlation between all pairs of stochastic variables can be estimated from the variation on the level of the individual samples. The diagonal of **E** contains the error terms associated with all the stochastic variables.

The inaccuracy in the estimate of average landing weight of a category $\overline{w}_c$ is simply the standard error of the mean, and can be calculated as the square root of the sample variance divided by the number of samples in that category S_c.

The inaccuracy in the estimate of the number of fish per category summed over all sampled landings again must be approximated by application of the error propagation formula. If the possible correlation between sample proportions from various landings can be ignored, the variance of the estimate is

$$[7] \qquad VAR(n_c) = \sum_s \left(\frac{n(s)_{c,s}^{\;2}}{\hat{p}(s)_{c,s}^{\;4}} VAR\left(P(s)_{c,s}\right) \right)$$

Here, $\hat{p}(s)_{c,s}$ denotes the observed sample proportion per category per sampled landing. Estimation of the variance of the sample proportion is problematic, as it depends on the true probability of fish being sampled. For the moment I will use an ad-hoc formula, with the number of fish sampled as denominator, so the estimate of the variance becomes $\hat{p}(s)_{c,s}\left(1-\hat{p}(s)_{c,s}\right)/\,n(s)_{c,s}$. Although incorrect, it won't be far off.

Finally, the inaccuracy in the estimate of the probability that fish are of a particular age given market category $P_{a,c}$ is defined if $P_{a,l}$ and $P_{l,c}$ are independent random variables. The variance of the estimate is

$$[8] \qquad VAR\left(P_{a,c}\right)=\sum\nolimits_{l}\left(\hat{p}_{a,l}{}^{2}\,VAR\left(P_{l,c}\right)+\hat{p}_{l,c}{}^{2}\,VAR\left(P_{a,l}\right)+VAR\left(P_{l,c}\right)VAR\left(P_{a,l}\right)\right)$$

Here, $\hat{p}_{a,l}$ denotes the observed proportion of fish of a particular age within a length class, and $\hat{p}_{l,c}$ denotes the observed proportion of fish of a particular length class within a market category. The variances depend on the true probabilities $P_{l,c}$ and $P_{a,l}$, respectively, but may be reasonably approximated, provided the sample is sufficiently large, by inserting the sample proportions to obtain $\hat{p}\left(1-\hat{p}\right)/\,n$. A problem with this approximation is that the number of fish sampled within each length class is rather small, in any case less than the number of fish sampled per market category.