Proportions-at-weight, a new Observation type

Matt Dunn, Ian Doonan, Teresa A'mar

February 2021

1. Overview

Overview

2. Conceptual Description

The observation process_removals_by_weight was added to allow the use of fish market data, where fish weights have been measured instead of fish lengths or ages, which are collated into proportions-at-weight. The observation is strictly from retained catch and is therefore associated with mortality from a defined fishery. When discarding is included in the model, the retained selectivity should be used for these data. If there are no discarding, then the total catch selectivity is used. However, for current coding purposes, this observation must be linked to a fishery as defined in a mortality_Instantaneous block, which ignores discards and so the selectivity is that defined in the mortality_Instantaneous block.

The expected proportions-at-weight are derived from normalizing the expected numbers-at-weight. There are two stages in calculating the expected numbers-at-weight. Firstly, CASAL2 calculates the expected numbers-at-length from the numbers-at-age using the age-length and distribution in the `@age_length` block. Secondly, the numbers-at-length are converted into numbers-at-weight using the length-weight relationship in the `@length_weight` block and its distribution of weight about the mean weight-at-length. The length-weight distribution is currently only applied in CASAL2 in the `@observation` block for weight composition data and is therefore specified here, as the CV of either a normal or lognormal distribution.

The user must specify the units of weight for the proportions-at-weight observations (which may be g or kg), a vector containing the lower edge of each weight bin, and a vector containing the proportions in each weight bin. Note that units for weight in the proportions-at-weight do not have to be the same as specified in `@length-weight`. Observation-specific weight bins must be a sequential subset of the model weight bins, with no missing or added values. Observations may be specified for any category used in the partition, or for some combinations of them [LINK TO CATEGORY SECTION?], e.g., for both males and females separately, or alternately, one set for combined sex. The weight bins must be the same for each year; if this is not the case, then years for which they are different need to be entered as different process_removals_by_weight blocks.

If there is no plus group, i.e., weight_plus=false, then CASAL2 requires a vector of proportions-at-weight of length n+1, where n is the number of lower edges of each weight bin supplied (the final value provides the upper limit to the final bin). If weight_plus=true then CASAL2 expects a vector of proportions-at-weight of length n. The last proportion represents the numbers from the last length bin to the maximum weight the age-weight relationship allows [not sure how we should specify this?].

CASAL2 generates a warning if the mean weight estimated for the youngest age in the partition is greater than the lower size of the first weight bin. This is to guard against including weight observations that may have a substantial contribution from fish younger than the youngest age in the partition.

The only likelihood currently available in CASAL2 for proportions_at_weight observations is the multinomial, with effective sample sizes for each year provided as error_values. Note that in the implementation of the multinomial likelihood in CASAL2, the weight bins having a value of zero will have no contribution to the likelihood.

No specification is made for the specific time within the timestep that the same was taken. This is because the sample is linked to a fishery, where removed are defined to be at the mid-point.

Proportions-at-weight processes:

- process_removals_by_weightfishery using Instantaneous_Mortality;

- process_removals_by_weight_retainedfishery using Instantaneous_Mortality_Retained; LEAVE OUT FOR NOW IJD?

- process_removals_by_weight_retained_totalfishery using Instantaneous_Mortality_Retained LEAVE OUT FOR NOW IJD?

documented in Section 7.1.3 (of the User Manual?). Specific process observations (i.e., fishery tied observations; obs fitted to the middle of the fishery since it represents it)

We have "general" obs like process_proportions_at_length, which are used for surveys, but also for fisheries; they can be fitted into any place throughout the time-step. Weight frequencies are not collected on surveys, so we do not need this observation yet.

Length composition observation types for fisheries.

Linked to mortality_instantaneous, CASAL2 has these fishery length composition data observations:

- process_removals_by_length: works with process of type mortality_instantaneous;

- process_removals_by_length_retained: works with process of type mortality_instantaneous _retained; and

- process_removals_by_length_retained_total: works with process of type mortality_instantaneous _retained

Fisheries process: mortality_instantaneous

Supply time step.

Specific fishery is a method in the table method in the mortality_instantaneous block. For one stock, it only makes sense to have one mortality_instantaneous block that covers all fisheries on the stock (whatever time-step they occur in) and specifies the proportions of M that occurs in each time-step. Rather than covering one process in a specific time-step, mortality_instantaneous blocks cover the full year which makes it different from other processes. mortality_instantaneous blocks need to be like this so that M and F can occur simultaneously and to also make sure the various U_max parameters are evaluated at the same time if fisheries co-occur.

Alt fisheries process mortality_instantaneous _retained: same as above, but with discards and retained catch

Copied from the manual, but changed into proportions-at-weight.

Proportions-at-weight observations can be supplied as

- a set of proportions for a single category (see example);

- a set of proportions for multiple categories; or

- a set of proportions across aggregated categories.

The method of evaluating expectations are the same for all three types of proportions.

Defining an observation for multiple categories extends the single category observation definition. It is used to model a set of proportions over several categories by weight bin. For example, to specify that the observations are of the proportions of male or females within each weight bin, then the subcommand categories is

```
categories male female
```

The vector of proportions will have the proportion for males over the specified weight bins, followed by the proportions for females. The sum over male and female proportions should be 1 (i.e., it implicitly has a sex ratio).

Defining an observation across aggregated categories allows categories to be aggregated before the proportions are calculated. To indicate that two (or more) categories are to be aggregated, separate them with a "+" symbol. For example, to specify that the observations are of the proportions of male and females combined within each weight bin, then the subcommand categories is

```
categories male + female
```

CASAL2 then requires that there will be a single vector of proportions supplied, with one proportion for each weight bin, and that these proportions sum to one.

The latter form can then be extended to include multiple categories, or multiple aggregated categories, e.g., for an east and west combined sex proportions that incorporates an area ratio we could use

```
categories male.east + female.east male.west + female.west
```

CASAL2 then requires that there will be a vector made up by a concatenated of proportions for east and another for west, and that these proportions sum to one.

3. Technical Description

The observation is supplied for a given year and time-step, for some selected age classes of the population (i.e., for a range of ages multiplied by a selectivity that is associated with the process). For length selectivities in an age-based model, we must apply the selectivity onto the age-length matrix, so do we generate the length-age matrix first, then apply whatever selectivity is needed?

The expectations from this observation are generated whilst the process is being executed. There is a chain of expectations starting with that for ages, then converting this into an expectation for lengths, and lastly, converting lengths into expectations over weight bins.

The expectation of numbers at age a for category c from exploitation method m ($E[N_{a,c,m}]$) are

$$E[N_{a,c,m}] = N_{a,c}U_{a,m}S_{a,c,m}0.5M_{a,c} \tag{3.1}$$

where $N_{a,c}$ are the numbers at age in category c before the process is executed, $U_{a,m}$ is the exploitation rate for age a from method m, $S_{a,c,m}$ is the selectivity, and M is the natural mortality. This is OK for age-based selectivities, but for length based selectivity (in an age-based model), a form of age to length transition is done, but the probabilities use numerical integration within an age bin, so there may be speed advantages to incorporating this into the age-to-length transition matrix calculations below. Given $U_{a,m}$ needs $S_{a,c,m}$, it is cleaner to keep length-based selectivity calculations separate for now.

3.1. Part 1: age to length transition

We now use the specified length bins in the `@model` command block to generate expectations. This is a transition matrix,Tal,that converts proportions of an age into the vector of lengths bins; columns are ages, rows are length bin index so $Nl_{c,m} = \mathbf{Tal}N_{c,m}$, where Nl and N are (column) vectors.

We drop then m subscribe (method) for here on.

For the normal distribution of length about an age, elements of **Tal** for length bin l and age a in category c are given by $\zeta_{l,a,c} = \Phi\left(\frac{L_{l+1}-\bar{L}_{c,a}}{\sigma_{c,a}}\right) - \Phi\left(\frac{L_l-\bar{L}_{c,a}}{\sigma_{c,a}}\right)$, where Φ is the standard normal cumulative function with mean length $\bar{L}_{c,a}$ and standard deviation $\sigma_{c,a}$. For a log-normal distribution, the $L_l, \bar{L}_{c,a}$ and $\sigma_{c,a}$ must be on the log scale.

Notice that when the length distribution has a probability of lengths below L_1 these fish are "lost" which is different to SS3 manual equation A.1.14 where fish in bin 1 are for fish below L_1.

The calculation for last length bin, l_{max}, depends on whether a length plus-group is defined in `@model` length bins. For no plus-group, l_{max} is one less than the number of lengths specified since the last length caps the interval for bin l_{max}. If we index the last length as l_{cap} then

$\zeta_{l_{max},a,c} = \Phi\left(\frac{L_{l_{cap}} - \bar{L}_{c,a}}{\sigma_{c,a}}\right) - \Phi\left(\frac{L_{l_{max}} - \bar{L}_{c,a}}{\sigma_{c,a}}\right).$

For a plus-group in `@model` length bins, l_{max} indexes the bin that starts at the last specified length, but includes all length larger than it. $\zeta_{l_{max},a,c}$ is now $1 - \Phi\left(\frac{L_{l_{max}} - \bar{L}_{c,a}}{\sigma_{c,a}}\right)$.

If growth is not time varying, then **Tal** should be calculated once. If length-weight relationships are time-varying, then **Tal** should be saved and re-used.

Warnings

The age ranges specified in `@model` can exclude the youngest ages, but invariably it has a maximum age, usually with a plus-group. Both situations can create bias in predicting length distributions in certain circumstances. If the age plus-group is not located where the mean length is at L_∞, some parts of the length distribution within the age plus-group are not represented and this will result in a bias for length composition data where a length plus-group is specified. This is only an issue in the early part of the model run when the fishdown is occurring, after which it will disappear as the age plus-group is emptied. A bias may also occur when the youngest ages are excluded if their length distribution(s) overlaps with the smaller length bins in the model. We should detect this, if possible, and issue a warning to alert the practitioner. Can we make growth functions take ages outside the age range to predict mean length (it may do this already - TA)? If so, CASAL2 can calculate how much the age class below the minimum age (if this is greater than one) contributes to the `@observation` length composition.

For age plus-groups, CASAL2 should be able to generate the cumulative distribution for the terminal length distribution and for years

```
max_age plus c(5, 10, ...)
```

which can be inspected by the user to check the likely consequences of their proposed configuration. For age plus-groups, what age to use for the terminal length distribution is unclear, maybe just `age.max`. In CASAL, age plus-groups have a parameter to specify the length to used for the weight calculations, which implies an age. In CASAL2, I am unsure what happens given the manual entries:

```
@model
length_plus_group The mean length of length plus group #can this be specified statically (IJD)
default 0  #??

age_plus The oldest age or extra length midpoint ,plus group size, as a plus group??█
Type: boolean
Default: true
Value: true, false
```

I am unsure what has been done here and what would be better (IJD). For age plus-groups, should we have `@model/age_plus_age/default age_max` to define the terminal length distribution and mean length for weight calculations? Similarly for length plus-groups, should we have

`@model/length_plus_length/default length_max` to define the terminal weight distribution and mean weight for weight calculations? Maybe the `@model/length_plus_group` above does this already?

3.2. Part 2: age to weight transition

Here, **Tal** is converted into **Taw**, the age to weight transition matrix. First, a version is calculated using the weight bins specified in `@model`, **TawM**. Second, this is transformed into the specific weight bins specified in the `@observation` section, **TawO**.

The `@observation` weight bins must be consistent with the `@model` bin boundaries, i.e., `@observation` bins cannot cut a `@model` bin into two or extend outside its range. It can concatenate `@model` bins and it does not have to cover the full range. `@observation` bins cannot have "holes" in it. Holes and bin consecutive order in the specifications should be checked for in both the `@observation` and `@model` versions (this has caused problems for length bins).

An intermediate matrix is needed that converts length into a weight distribution, **Tlw**. This uses the `@model` length bins so it fits to the **Tal**, i.e. **TawM** $= mathbfTalmathbfTlw$.

For the normal distribution of weight about length, elements of **Tlw** for weight bin w and length l in category c are given by $\zeta_{w,l,c} = \Phi\left(\frac{W_{w+1}-\bar{w}_{c,l}}{\sigma_{c,l}}\right) - \Phi\left(\frac{W_{w}-\bar{W}_{c,l}}{\sigma_{c,l}}\right)$, where Φ is the standard normal cumulative function with mean weight $\bar{W}_{c,l}$ and standard deviation $\sigma_{c,l}$. Is there going to be a log-normal version (MD)?

We are not allowing a weight plus-group (?) so the above defines calculations needed. A plus-group can be made by increasing the $W_{w_{max}}$ to a unlikely value.

Now **TawM** $=$ **Tal** $\% * \%$ **Tlw**.

Do we need the `@model` weight bin? If there are more than one series of weight composition data then `@model` weight bins will save some time. Is this the case (MD)? The Φ calculations are expensive since they use exp() and raising to powers up to 5 (or is it 7?) so calculating them once could be advantageous.

To get the finial transition matrix, **TawO** we may need to delete `@model` bins (columns) from both ends and potentially add over internal bins. If wm_i are the lower weight bin boundaries (`model`) for $i = 1...m-1$ bins (m^{th} value defines the cap for the last bin), and wo_j are the lower boundary values (`observation`) for $j = 1...n-1$ bins, then an j index can be assigned to each `model` bin based on $wo_j <= wm_i < wo_{j+1}$; for no fits than 0 is assigned (i.e., deletion). R code is

```
j.wm #j index for model bin into observation bin
TawO <- matrix(rep(0,(n-1)*(max.age - min.age +1)),ncol=m-1)

for(i in c(1,m-1)){
     TawO[,j.wn[i]] <- TawO[,j.wn[i]] + TawM[,i]
     }
```

If growth and length-weight relationships are both not time-varying, then **TawO** should be saved and re-used for the other years. If both growth and length-weight parameters are not being estimated, and both are not time-varying, **TawO** only needs to be calculated once.

3.3. Predicted weight composition

$\overrightarrow{N}_{c,m}$ is the n_{age} x 1 column vector of expected numbers-at-age based on $E[N_{a,c,m}]$ from above. Doing this isolates the transition matrix from selectivity in case it is time-varying.

First we calculated the predicted weight composition for single fish measurements (i.e., the above derivation) by $\overrightarrow{W}_{c,m} = \mathbf{TawO^{T}_{c,m}} \; \overrightarrow{N}_{c,m}$. [$nx1$ vector $n \ x \ n_{age}$ matrix times $n_{age} \ x$ 1 vector]. Then normalize $\overrightarrow{W}_{c,m}$.

3.4. Accounting for data being mean weight composition

$\overrightarrow{W}_{c,m}$ is put into the multinomial log-likelihood.

Do we adjust sample size by the mean number of fish in a box? Or, do we allow reweighting to do the adjusting?

DOES NOT ACCOUNT FOR DATA BEING MEANS OVER n_i FISH IN EACH BOX.

The data are for mean weights for individuals about the same size (=length?) (yes?). So we can use the length-weight relationship to predict its mean weight. The length-weight has a cv (say 10%) which reflects the variation of weight for a length, but this is for one fish, not the 4 that are in the box. If all boxes have 4 fish the we just use 10/2 = 5% for the cv = easy fix.

But, does the mean number of fish vary significantly across the boxes? Numbers from 3 to 5 could be approx. by 4. With the switch to juveniles, does the number/box go to 10? 20?

The later might need variable adjustments to the cv across the weight bins, i.e., large numbers decreasing down to 4. In that case, we will need a mean number/box for each bin and do the adjustments when calculating transition matrix, i.e., cv_bin = cv/sqrt(n_bin).

For the time being, we can leave it as is, i.e., adjust weight distribution cv to be cv/2.

Warnings

There might be the same issues as for lengths, except that we will not allow a weight plus-group

4. Proposed Input Format

Example input block

```
@observation Observed_weight_frequency_east
type process_removals_by_weight
method_of_removal EastChathamRise  # fishery
time_step Summer # Not truly needed, but length code needs it currently
mortality_instantaneous_process instant_mort
length_weight_cv 0.1
length_weight_dist lognormal
years 1991 1992
categories male
weight_unit kg
delta 1e-5  # robustification value for the likelihood; default 1e-11
likelihood multinomial
weight_plus false        #not allowed?
weight_bins 1.8 1.9 2.0 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3.0 3.1 3.2 3.3 3.4 3.5 3.6 3.7 %

table obs
1991 0.002 0.010 0.041 0.094 0.126 0.086 0.073 0.051 0.045 0.050 0.037 0.025 0.016 0.017 0.011
1992 0.002 0.006 0.016 0.018 0.027 0.038 0.058 0.069 0.051 0.061 0.063 0.052 0.032 0.028 0.021
end_table
```

```
table error_values
1991 25
1992 25
end_table
```

5. Test Plan

Simulate some test data in R which can become a unit test.

6. Text to be added to the User Manual

observations 7.7x

Process removals by weight

Removals by weight observations are observations of the relative number of individuals at weight, part way through a process of type `mortality_instantaneous`. This observation is exclusively associated with the process of type `mortality_instantaneous`, and will produce an error if associated with any other process type.

The observation is supplied for a given year and time-step, for some selected age classes of the population (i.e., for a range of ages multiplied by a selectivity that is associated with the process).

The expectations from this observation are generated whilst the process is being executed. The expectation of numbers at age a for category c from exploitation method m ($E[N_{a,c,m}]$) are

$$E[N_{a,c,m}] = N_{a,c}U_{a,m}S_{a,c,m}0.5M_{a,c} \tag{6.1}$$

where $N_{a,c}$ are the numbers at age in category c before the process is executed, $U_{a,m}$ is the exploitation rate for age a from method m, $S_{a,c,m}$ is the selectivity, and M is the natural mortality.

The observation class accesses the variable $E[N_{a,c,m}]$ from the process and applies the age-length relationship specified in the model. This converts numbers-at-age to numbers-at-age and -length. In turn, numbers-at-length and -age are converted into numbers at -age and -weight using the length weight relationship and the spread of weights about a length, which are then converted to numbers-at-weight. The observations are aggregated by method and category depending on how the user specifies the observation, before converting numbers-at-weight to proportions and calculating the likelihood.

Similar to the proportions-at-length and removels-by-weight observation types, the user must supply a vector of weight bins. The observation-specific weight bins must be a sequential subset of the model weight bins, with no missing or added values. For example, if the model weight bins are `0 5 10 15 20 25 ... 100`, then the observation-specific weight bins can be `20 25 30 35 40 45 50` but not `20 30 40 50`. The length bins used int eh intermediate step to get the numbers-at-age and -length uses the model length bins. For time invarient growth and length weight relationships, the transition matrix from numbers-at-age to numbers at weight is calulated once at the start of a run and stored to be reused in other years to speed up htis observation class.

Currently, CASAL2 does not allow a weight plus group. WARNING: The user should check that the weight range for fish in the age plus-group is consistent with the specified weight bins for the years with weight compositional data (e.g., weight compositional data collected when the stock has been fished down and so the age plus-group has insignificant fish in it meets this criteria). SImarly for model length bins.

```
@observation observation_fishery_WF
type process_removals_by_weight
...
years  1993 1994 1995
method_of_removal FishingEast
mortality_instantaneous_process instant_mort
weight_bins 0 20 40 60 80 110
delta 1e-5
table obs
1993    0.0   0.05    0.05    0.10    0.80
1994    0.05  0.1     0.05    0.05    0.75
1995    0.3   0.4     0.2     0.05    0.05
end_table

table error_values
1993 31
1994 34
1995 22
end_table
```

Likelihoods that are available for this observation are the mulitnomial and the lognormal. See Section **??** for information on the likelihoods.

7. References

X