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. We do not know of discard data being collected for fisheries that provide weight composition data (currently in 2021, just the bluenose fishery in New Zealand) and so this feature is being proved in the only 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, a vector containing the proportions in each weight bin, and a vector containing the mean number of fish in a box at that weight bin (to adjust the cv of the length-weight distribution to account for the fact that these data are mean weights, not weights for a single fish as found in a length composition). 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 in a sequence; there is no `@model/weight_bins` as in length composition observations. 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 a different process_removals_by_weight blocks.

CASAL2 does not allow a weight plus group, so the data are specified in 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). Note that a plus group can be mimicked by using a large terminal weight bin value.

CASAL2 generates a warning if the mean weight estimated for the youngest age in the partition is smaller 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 sample was taken. This is because the sample is linked to a fishery, where removals and observations are defined to be at the

mid-point.

Proportions-at-weight processes:

- process_removals_by_weight_fishery using Instantaneous_Mortality;
- process_removals_by_weight_retained_fishery using Instantaneous_Mortality_Retained; Not coded in this release.
- process_removals_by_weight_retained_total_fishery using Instantaneous_Mortality_Retained; Not coded in this release.

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 currently do not need this observation.

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

Weight data is currently collected in processing sheds and so discarded data is not available. Discard data is not usually collected by observers on fishing vessels and so, again, there is generally no discard data. Hence, only the weight equivalent of process_removals_by_length is needed currently.

Fisheries process: mortality_instantaneous

Supply time step.

A specific fishery is a `method` in the `table method` sub command 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.

Proportions-at-weight observations 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). [CASAL2 will issue a warning and normalize observation vectors that do not sum to 1. For length observations, the vector must sum to 1 otherwise an error is logged which seems to be unhelpful since generating these vectors may not sum to 1 if rounded, and requires a complete re-calculation if end parts are to be excluded.]

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 sub-command 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. CASAL2 will issue a warning and normalize observation vectors that do not sum to 1.

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. CASAL2 will issue a warning and normalize observation vectors that do not sum to 1.

3. Technical Description

3.1. Input sub-commands for weight frequency observation block

A weight composition data series is signaled in the input files by:

```
@obervation <label to be supplied)
type process_removals_by_weight
<other sub-commands>
```

· ⋮

The sub-commands that are available are shown in Table 3.1.

3.2. Numbers-at-age Expectation

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. 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.

Table 3.1: Input sub-commands for weight frequency observation block. "<X>" denote values or vectors that the user must supply (without the <>).

type process_removals_by_weight	Observation is for weight frequency data
method_of_removal <fishery label>	Fishery that this observation belongs to
time_step <time step>	Not truly needed, but length code needs it currently
mortality_instantaneous_process <mortality block label>	Block that the fishery is defined in
terminal_length <length>	Length cap for length bins to use in the age-length transition matrix. Length bins used are: 1, 2, 3, ...(<length> - 1). Default is $L_\infty + 4 * cv * L_\infty$. If different growths are used over the categories this observation covers, the growth with the largest L_∞ is used.
length_weight_cv <cv>	Cv for weight-at-length distribution
length_weight_dist <distribution label>	Weight-at-length distribution: either lognormal or normal
years <vector of years>	Specify years that have data.
categories <category specification>	Categories that the observation was collected from.
weight_unit <unit label>	Either kg, tonnes, or gm?
delta <value>	Specify the robustification value for the likelihood; default 1e-11.
likelihood <likelihood label>	Distribution for the weight frequency. Only option is currently multinomial (N scales the variance .
weight_bins <vector of lower bound for each bin> <cap for last bin>	No plus-groups allowed.
length_bins_n <vector of n, one for each length bin >	Cv for the weight-at-length distribution in $i^(th)$ bin is length_weight_cv $/\sqrt{n_i}$.
fishbox_weight <value>	alt approach for length_bins_n
fishbox_length_weight <label to a length-weight block>	al^b to use
table obs ... end_table	Table sub-commands that enclose observations. Each row had the year followed by a vector of the fractions for each weight_bin. The data should sum to one, but if not, CASAL2 will normalise it with a warning.
table error_values ... end_table	Table sub-commands that enclose the observation Ns. Each row has the year and the N for that observation.

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.

3.3. Part 1: age to length transition

This observation uses its own specified length bins to make sure the intermediate age-length transition is properly specified. This is in contrast to length compositions, which uses both an observation `length_bins` and a `@model/length_bin` specifications, i.e., specifications of weight composition data are independent of specification for length data. Here, the length bins (lower bound) used are from 1 (unit as specified in `@age_length`) in steps of 1 unit up to the terminal length-1. The terminal length caps the last length bin, i.e., plus groups are not allowed. The default value of the terminal length is based on 4 standard deviations above L_{inf}. If there are different candidate growth curves for the categories used in the observation, then the growth with the maximum L_∞ over these categories is used. To override the default, use sub-command `terminal_length <value>` in the observation block.

The expected length distribution is calculated via a transition matrix,**Tal**, that converts an age into the vector of proportions by lengths bins; columns are ages, rows are length bin index so $Nl_{c,m} =$ **Tal**$N_{c,m}$, where Nl and N are (column) vectors indexed by category (c) and method (m), i.e., $Nl_{c,m,l} = \sum_{a=A_{min}}^{A_{max}} \mathbf{Tal}_{l,a,c}N_{c,m,a}$, where A_{min} and A_{max} are the specified minimum and maximum ages in `@model` block.

We drop the m subscribe (method) from here on since it is clear which method is meant (i.e., data is from one fishery (=method in the `instantaneous_mortality` block). Categories index (c) is still relevant.

For the normal distribution of length given 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 $\sigma_{c,a} = cv_{c,a}\bar{L}_{c,a}$; $cv_{c,a}$ is specified in the relevant `@age_length` block. For a log-normal distribution, the $L_l, \bar{L}_{c,a}$ and $\sigma_{c,a}$ must be on the log scale, i.e.,

$$\zeta_{l,a,c} = \Phi\left(\frac{\log L_{l+1}-\left(\log\bar{L}_{c,a}-\sigma^2_{c,a}/2\right)}{\sigma_{c,a}}\right) - \Phi\left(\frac{\log L_l-\left(\log\bar{L}_{c,a}-\sigma^2_{c,a}/2\right)}{\sigma_{c,a}}\right)$$

where $\sigma_{c,a} = \sqrt{\log(cv_{c,a})^2+1}$, i.e., log scale standard deviation.

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.

Recommendations

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 recommend that the model starts at age 1 and finishes at a sufficiently old age that the mean length is almost at L_∞ (especially since the need to save memory space and speed up the program are not such a consideration these days).

As a check, CASAL2 will output the predicted mean length and the length at 3 x standard deviation above the mean for ages at `max_age`, $max\_age + 5$ and $max\_age + 10$ years. This will mean that growth functions can take ages outside the `@model` age range to predict mean length (it may do this already?).

3.4. Part 2: length to weight transition leaading to age to weight transition

Here, **Tal** is converted into **TawO**, the age to weight transition matrix using the weight bins specified in the observation block, i.e., we do not have a `@model/weight_bins` sub-command that length data has. The specific weight bins are specified in the `@observation/weight_bins` sub-command. Weight bins must be in consecutive order and CASAL2 should check that this is so.

An intermediate matrix is needed that converts length into a weight distribution, **Tlw**. This uses the generated length bins based on `@observation/terminal_length` sub-command so it fits to the columns in **Tal**, i.e., **TawO** = **Tal** %*% **Tlw**.

The numbers-at-weight-bin (column vector) is given by $NW_{c,w} = \sum_{l=L_{min}}^{L_{max}} \mathbf{Tlw}_{w,l,c} Nl_{c,l}$, where $L_{min} = 1$ and $L_{max} =$ (terminal length-1) . From the section above, the numbers-at-length vector is $Nl_{c,l} = \sum_{a=A_{min}}^{A_{max}} \mathbf{Tal}_{l,a,c} N_{c,m,a}$. Substituting for $Nl_{c,l}$ and summing over the length bins

$NW_{c,w} = \sum_{l=L_{min}}^{L_{max}} \mathbf{Tlw}_{w,l,c} \sum_{a=A_{min}}^{A_{max}} \mathbf{Tal}_{l,a,c} N_{c,m,a}$ which rearranges to

$NW_{c,w} = \sum_{a=A_{min}}^{A_{max}} \left(\sum_{l=L_{min}}^{L_{max}} \mathbf{Tlw}_{w,l,c} \mathbf{Tal}_{l,a,c} \right) N_{c,m,a}$.

For the normal distribution of weight about length, elements of **Tlw** for weight bin w and length l in category c are given by $\zeta_{w,l,c} = \Phi\left(\frac{\bar{W}_{w+1} - \bar{W}_{c,l}}{\sigma_{c,l}}\right) - \Phi\left(\frac{\bar{W}_{w} - \bar{W}_{c,l}}{\sigma_{c,l}}\right)$, where Φ is the standard normal cumulative function with mean weight $\bar{W}_{c,l}$ and standard deviation $\sigma_{c,l} = cv_{weight} \bar{W}_{c,l}$; cv_{weight} is specified as a constant for all lengths and categories in the `@observation/length_weight_cv` sub-command, but the l index is needed to account for more than one fish per fish box (see below). Again, there can be a log-normal version, i.e.,

$$\zeta_{l,a,c} = \Phi\left(\frac{\log \bar{W}_{w+1} - \left(\log \bar{W}_{c,l} - \sigma_{c,l}^2/2\right)}{\sigma_{c,l}}\right) - \Phi\left(\frac{\log W_{w} - \left(\log \bar{W}_{c,l} - \sigma_{c,l}^2/2\right)}{\sigma_{c,l}}\right)$$

where $\sigma_{c,l} = \sqrt{\log(cv_{c,l})^2 + 1}$, i.e., log scale standard deviation.

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

Now the age-weight transition matrix, **TawO** is given by **Tal** %$*$% **Tlw**.

$\mathbf{TawO}_{w,a} = \left(\sum_{l=L_{min}}^{L_{max}} \mathbf{Tlw}_{w,l,c} \mathbf{Tal}_{l,a,c} \right)$.

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.

3.5. Predicted weight composition for single fish measurements

$\overrightarrow{N}_c$ 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 $\overrightarrow{W}_c = \mathbf{TawO}_\mathbf{c}^\mathbf{T} \overrightarrow{N}_c$. [$n_{weight}$x1 vector n_{weight} x n_{age} matrix times n_{age} x 1 vector].

Element-wise, $\overrightarrow{W}_{c,w} = \sum_{a=A_{min}}^{A_{max}} \mathbf{TawO}_{w,a,c} \overrightarrow{N}_{c,a}$. Then normalize $\overrightarrow{W}_c$.

$\overrightarrow{W}_c$ is put into the multinomial log-likelihood.

3.6. Accounting for data being mean weight composition

The above section does not account for data being means over n_i fish in each fish box, where i indexes length bins and since fish length in a particular fish box are similar, we can use the length bin boundary to approximate length. For the bluenose data, n_i can range from 2 to 18, with a median of 4. Using the length-weight relationship to predict its mean weight over several fish, we need to adjust the length-weight cv (say 10%) to reflects the variation of weight for several fish which are likely to be at similar lengths, i.e., we should us $cv_{weight}/\sqrt{n_i}$. If all boxes have 4 fish the we just use 10/2 = 5% for the cv.

To accommodate this structure, the user needs to supply a vector with a n_i for each length bin, using the `length_bin_n` sub-command. This is not strictly correct, but it should be good enough. The re-weighting R code needs to be adjusted to take this structure (I think), which may provide further ways to adjusted the error structure.

The cv_l in $\sigma_{c,l} = cv_l \bar{W}_{c,l}$ above is a constant over both categories and lengths and so reflects the distribution of a single fish for the same number of fish in each fish box.

cv_l needs setting to $cv_{weight}\sqrt{n_l}$. How best to do this at set up? We could use the `length_bin_n` sub-command, but the number of length bins may be hard to calculate when using the default terminal length. Or, do we calculate n_l from $w_{fishbox}/(al^b)$ rounded into an integer (using some mean fishbox weight as they are likly to be similar)? The later will require another sub-command like `fishbox_weight <value>`. Or allow both so that if `fishbox_weight` is present, $w_{fishbox}/(al^b)$ is used. However, a and b can vary by category so we will need a `fishbox_length_weight <label for a length weight block>` to pick out one value set. «««««««««««««««««««<

Recommendations

There might be the same issues as for lengths, except that we will not allow a weight plus-group. Should we try to mean weight and mean + 4 * standard deviation weight at the same ages as for lengths as a check that the weight bins are sufficient for the age plus group, but also ages 1 and 2, say, for the lower end (now mean - 4 * standard deviation weight)?

Utility R functions

Need check extract function from the fit output works for weight frequencies. Otherwise, need to write or adapt some code. We probably need some C++ report code.

4. Proposed Input Format

specify units?

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
```

```
terminal_length  25
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 #there are no others
#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 %
weight_bins_n   4    #expanded into a vector with all 4s in it

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