fish5103growth Modelling length at age and length distributions

Gunnar Stefansson and Lorna Taylor

19. desember 2016

Copyright This work is licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit http://creativecommons.org/licenses/by-sa/1.0/ or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

Acknowledgements
MareFrame is a EC-funded RTD project which seeks to remove the barriers preventing more widespread use of the ecosystem-based approach to fisheries management.
`http://mareframe-fp7.org`
This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no.613571.
`http://mareframe-fp7.org`
Háskóli Íslands
`http://www.hi.is/`

Efnisyfirlit

1 Lack of age data - background

1.1 Poor data is no excuse

Lack of age readings does not change the issue:

- The population dynamics are the same
- The problem simply becomes harder
- Poor methods result in poor data
- Need better methods with poor data
- Should attempt to get better data

1.1.1 Details

It is a common misunderstanding that simple methods of analysis are always appropriate when a researcher has “simple” data. It is, of course, true that a regression analysis using few or poor data points will lead to only a few variables appearing significant and this may be the reason for the misunderstanding. In fishery science, however, it is the population dynamics which are important and the *nature of population dynamics do not become simpler just because data are not available*.

Age data may not be available for a number of reasons, but mainly the reasons are either biological (there are no hard parts which can be used to identify annuli) or economical (it is too expensive to collect and analyse data).

The lack of age data does not change the primary issue when considering stock dynamics, which is to obtain an understanding of how the population will respond to different pressures. In particular, the lack of age data does not mean that there are no age groups in the population!

The lack of age data is not an excuse to use outdated methods or methods which are known to perform poorly. In particular, many "classical"methods have been extensively tested and found to give very unreliable estimates of population abundance and yield potential.

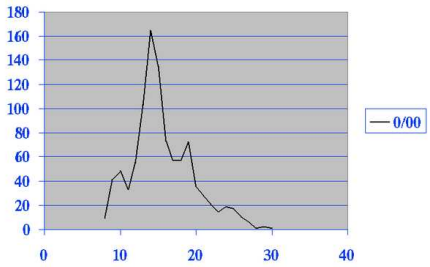
In many cases the lack of good data is really just a lack of organised collection and verification. In this case the single most important issue is really to organize a better data collection scheme.

Length data alone sometimes have enough resolution to give the same information as data sets with age readings. Unfortunately, the general rule is that it is more difficult to extract detailed information from poor data. It follows that in order to obtain sensible results one needs more elaborate and complex methods as the data gets poorer.

This implies the exact opposite to the popular view: Methods which work poorly on good data should never be used on poor data since then it is not even possible to see how poorly they perform. More elaborate methods which are based on valid statistical techniques are much more likely to provide appropriate results in general, also for simpler data sets.

1.2 Missing age readings

- Can often get growth from modes
- Can often use statistical methods to convert to age
- Can often use models of population dynamics and fit to length data
- Important: Can often get some (recruitment) data



Length distribution of Northern shrimp, *Pandalus borealis*, in parts per thousand (0/00) in each 1mm carapace length group, from a sample taken in Icelandic waters.

1.2.1 Details

In cases when age readings are not available, several approaches exist to stock assessments.

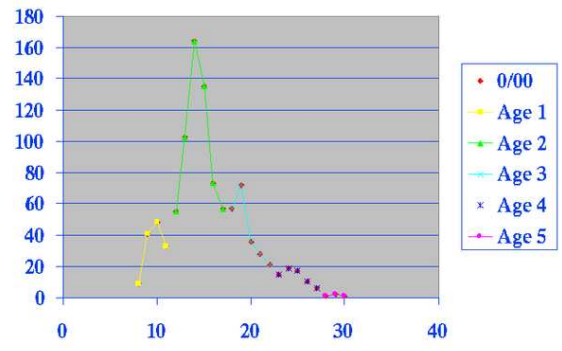
First, methods are available for converting length distributions to age compositions. In many cases these methods can be used to obtain reliable estimates of the age composition of catches.

Secondly, statistical methods exist for fitting age-based population dynamics models to data on length distributions. This topic is more than enough for a separate discussion and the remainder of this tutorial focuses on obtaining age compositions based on length measurements alone.

In the simplest examples, peaks in the length distribution are distinct enough to discern age groups. In particular it is often fairly easy to identify these cohorts visually (qualitatively) and this immediately gives some information on growth. However, more than growth data is needed for a complete understanding of the population dynamics of a species.

In many data sets it is possible to discern the recruitment part of the length distribution. This provides important information for many estimation methods.

1.3 Cohort slicing



A length distribution is sliced by assigning length groups surrounding a mode to an age group.

1.3.1 Details

The simplest method for converting length distributions to age compositions is probably cohort slicing. But, the method assumes that there is some prior knowledge of growth.

Note 1.1. Cohort slicing is conducted using data on length at age of different cohorts; the length distributions are simply sliced at the midpoints between the lengths at age to give an approximation of the number of fish in each age group.

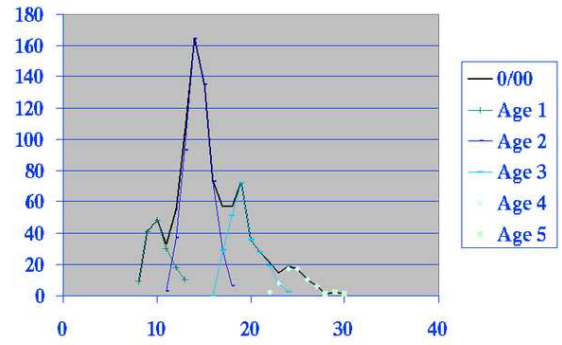
This method is clearly very simple in principle and when the peaks are reasonably clear, it is very easy to apply.

When the interest is only in obtaining the youngest one or two year-classes from the length distribution this method is more reliable than when used to slice the entire length distribution.

Detailed models of the population dynamics can include an internal model of length and age which can then be fitted to the sliced data without assuming that they come from only one age group.

Cohort-slicing is not to be recommended since better procedures exist.

1.4 Smoothing and interpolation



Simple visual smoothing methods can be used to guesstimate the fraction of each length group to go into each age group (not recommended).

1.4.1 Details

A slightly more sophisticated method than cohort slicing would use a simple interpolation mechanism to account for the fact that the in-between length groups should be allocated to more than one age group.

2 Growth models

2.1 Principles of mathematical modelling

A mathematical model of a biological process is, in its simplest form, just a formula used to describe the process.

Before attempting to fit complex statistical models, the procedure should be to envisage the biological processes, formulate them as mathematical models, and then study the behavior of the mathematical models.

Once the mathematical models appear to behave in accordance with the biological processes in question it is necessary to compare the models to data, which is where the statistics come in.

Part of the procedure is to plot and analyse data in order to verify which mathematical assumptions may reflect biological reality.

2.1.1 Details

A mathematical model of a biological process is, in its simplest form, just a formula used to describe the process.

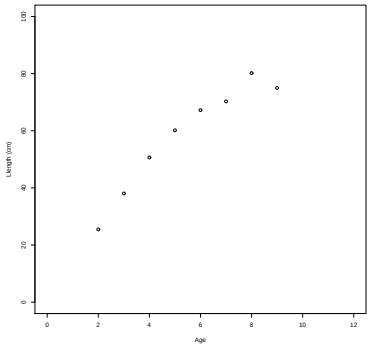
Before attempting to fit complex statistical models, the procedure should be to:

1. Envisage the biological processes
2. Formulate them as mathematical models
3. Study the behavior of the mathematical models

Once the mathematical models appear to behave in accordance with the biological processes in question it is necessary to compare the models to data, which is where the statistics come in.

2.2 Always first plot the data

First always plot the data - e.g. length against age.



2.2.1 Examples

Estimating a growth curve for a fish stock.

Example 2.1. First input some data and obtain estimates of mean length and numbers at each age - only use ages with minimal number of values.
http://tutor-web.net/fish/fish5103growth/lecture20/length-at-age.r

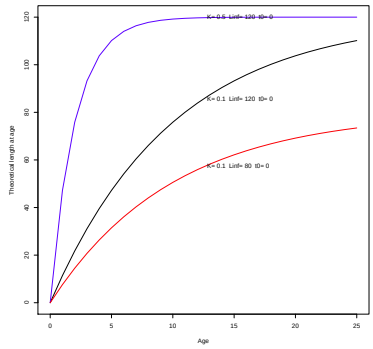
Example 2.2. Plotting mean length at age:
http://tutor-web.net/fish/fish5103growth/lecture20/mean-length-at-age.r

2.3 A model of fish growth

A model of growth - the von Bertalanffy growth equation

$$L_t = L_\infty \left(1 - e^{-Kt}\right)$$

$$L_t = L_\infty \left(1 - e^{-K(t-t_0)}\right)$$



2.3.1 Details

The von Bertalanffy model of growth describes the length at time (or age) t in terms of a few parameters. Though possibly not any more realistic, a parameter is often added in order to better fit data.

Definition 2.1. von Bertalanffy growth equation:

$$L_t = L_\infty \left(1 - e^{-Kt}\right)$$

$$L_t = L_\infty \left(1 - e^{-K(t-t_0)}\right)$$

This model only models a single trajectory. When used to model the growth of a group of fish it simply reflects the average length of the population.

In order to model a length distribution, this model needs to be extended to describe the distribution around the mean length at each age.

2.3.2 Examples

Example 2.3. If the age of a fish is stored in R as a vector, *a*, then one can use

```
lhat<-Linf*(1-exp(-K*(a-t0)))
```

to compute the predicted length at age.

It would be an improvement to implement this as a function...

Example 2.4. A more detailed examination of the behavior of the von Bertalanffy growth curve can easily be undertaken using R, e.g. through the following sequence of commands, or variations thereof:
http://tutor-web.net/fish/fish5103growth/lecture20/vonb-growth-curve.r

2.4 Mathematical models as functions in R

2.4.1 Details

Within R, more complex mathematical models are usually implemented as functions which take parameters as arguments and deliver fitted, or predicted values, as output.

2.4.2 Examples

Example 2.5. To take the von Bertalanfy growth function again, consider first a simple R function which predicts the length at age for a given set of parameters:

```
vonb<-function(Linf,K,t0){
  lhat<-Linf*(1-exp(-K*(a-t0)))
  return(lhat)
}
```

In many cases it is useful to store all the parameter values in a single vector:

```
vonb<-function(b){
  Linf<-b[1]
  K<-b[2]
  t0<-b[3]
  lhat<-Linf*(1-exp(-K*(a-t0)))
  return(lhat)
}
```

A better version is to include the age vector as an argument, here called "a":

```
vonb<-function(b,a){
  Linf<-b[1]
  K<-b[2]
  t0<-b[3]
  lhat<-Linf*(1-exp(-K*(a-t0)))
  return(lhat)
}
```

2.5 The sum of squares

Need to define the sum of squares deviations based on

$$y_i - \hat{y}_i$$

so use

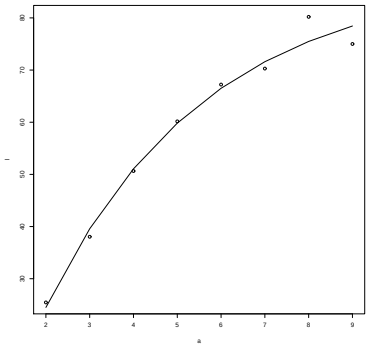
$$\sum_i (y_i - \hat{y}_i)^2$$

2.5.1 Examples

Example 2.6. Estimating a growth curve for a fish stock.
Define a new function which returns the sum of squares errors, for a given set of parameters. The data are assumed to be available in the directory where the function is called.

```
sse<-function(b){
 lhat<-vonb(b)
 s<-sum((l-lhat)^2)
 return(s)
}
```

2.6 Fitting a nonlinear growth model



2.6.1 Details

Nonlinear statistical models involve some nonlinear combinations of the parameters themselves (i.e. not the independent variables, so e.g. $y = \alpha + \beta x^2$ is in fact a linear model). Nonlinear estimation methods are therefore needed.

2.6.2 Examples

Example 2.7. Estimating a growth curve for a fish stock.
http://tutor-web.net/fish/fish5103growth/lecture20/fish-stock-growth-curve.r

Note that this did not take into account that there will be a difference in how accurate the various mean lengths at age are, though the initial selection process did limit the estimation to those ages with over 5 observations.

Example 2.8. A more complete example with a larger data set:
http://tutor-web.net/fish/fish5103growth/lecture20/fish-stock-growth-curve-expanded.r

3 Models of length distributions

3.1 Statistical and other models of length distributions

A fairly simple statistical model of length distributions is a combination of cohort length distributions, each of which is assumed to be for a specific probability density such as a Gaussian density.

The location of each density is centered on the mean length of the corresponding cohort with some standard deviation.

The multiplicative factors forming the combination reflect the relative strength of each cohort.

3.1.1 Details

Models of biological phenomena, such as length distributions, can be based on mathematical models of biological processes or simple statistical models which adequately describe the data at hand.

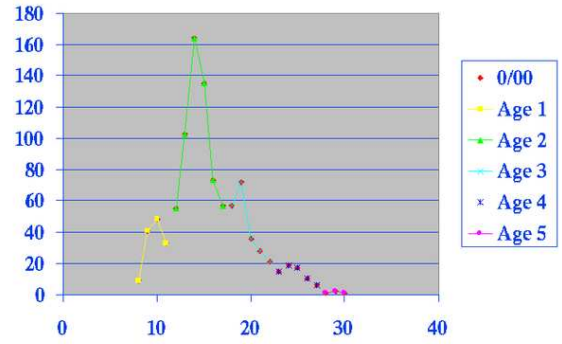
A fairly simple statistical model of length distributions is a combination of cohort length distributions, each of which is assumed to have a specific probability density such as a Gaussian density.

The location of each density is centered on the mean length of the corresponding cohort with some standard deviation.

The multiplicative factors forming the combination reflect the relative strength of each cohort.

In some cases discrete approximations to the Gaussian distribution are used whereas more commonly the cumulative probability is computed by integrating each length interval.

3.2 Cohort slicing



A length distribution is sliced by assigning length groups surrounding a mode to an age group.

3.2.1 Details

In some areas a method known as cohort slicing is commonly used to convert length distributions to age compositions. Remember, cohort slicing is the creation age compositions by taking the length at age of different cohorts and slicing the length distributions at their midpoints. This method is clearly very simple in principle and when the peaks are reasonably clear, it is very easy to apply.

Naturally, it is assumed that there is some information about growth.

In cases when the peaks are not clear, information from other sources may be used to decide on where to slice the length distribution. Thus, if tagging data are available, then they may indicate slicing points. Alternatively, in some years a clear cohort (large or small) may provide information as to which can be used for other years.

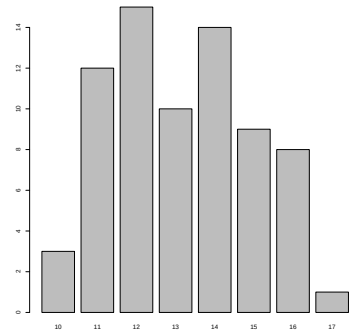
This methods has been tested quite extensively and found to be quite useful. Is it, however, clear that it suffers from several disadvantages. Obviously, the technique is not well founded biologically, statistically, or mathematically. More obviously, if a large year-class comes in, it will dominate several length groups, across the slicing points which may result in smearing.

Note 3.1. Smearing is where large year-classes become smaller and year-classes on both sides of the large year-class become larger.

The net result of smearing is that year-class variation will be underestimated and it may appear that recruitment is stable when in fact it is quite variable.

Thus, although this method is unquestionably useful, it does leave something to be desired in terms of how it behaves in various circumstances.

3.3 The distribution of length at age

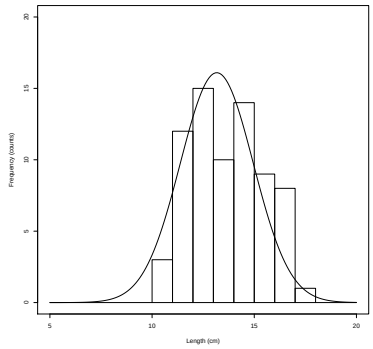


Some length measurements on age 1 cod in Icelandic waters in 1996.

3.3.1 Details

Within an age group fish will have different lengths. If a length distribution is to be modeled, this feature needs to be taken into account.

3.4 The Gaussian density and cdf



Length frequency and fitted Gaussian for a single age group.

3.4.1 Details

A probability distribution can be used to model the number of fish in each length group.

Start with a single age group and suppose the distribution of fish at length follows a Gaussian (normal) distribution. For simplicity, suppose the mean and standard deviation of length at age are known.

In this case the mean length and the standard deviations of the Gaussian distribution can be set to these known values. If the density is further multiplied by the numbers in the sample, a fitted curve ensues.

Although a Gaussian density can be fitted to the length data for a single age group, there is an inconsistency in this approach. Basically, the data correspond to aggregates since each length group consists of an interval. Typically, a length group l will actually contain all fish of length $l-\frac{1}{2}$ to $l+\frac{1}{2}$. The appropriate modeled probability should therefore reflect the area under the Gaussian density, between these two points.

Definition 3.1. Modeled probability under a Gaussian density for fish lengths between $l-\frac{1}{2}$ to $l+\frac{1}{2}$.

$$\Phi\left(\frac{(l+\frac{1}{2})-\mu_a}{\sigma_a}\right)-\Phi\left(\frac{(l-\frac{1}{2})-\mu_a}{\sigma_a}\right)$$

3.4.2 Examples

Example 3.1. The modeled probability reflecting the area under the Gaussian density can be done in R using:

```
pnorm((lgrp+0.5-mu)/sigma)-pnorm((lgrp-0.5-mu)/sigma)
```

where the mean length is *mu* and the standard deviation is *sigma*.

3.5 The proportion within a length group

3.5.1 Details

The density function of the Gaussian distribution with mean μ and variance σ^2 is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}$$

and the cumulative distribution is

$$F(x) = \int_{-\infty}^{x} \phi(t)dt = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Take a fixed age group of fish and assume that they are distributed along the length axis according to a Gaussian density, with some mean length (μ_a) and some standard deviation of length at age (σ_a). For this age group, the proportion of fish within length category l of width l is

$$\Phi\left(\frac{(l+\frac{1}{2})-\mu_a}{\sigma_a}\right) - \Phi\left(\frac{(l-\frac{1}{2})-\mu_a}{\sigma_a}\right)$$

since this is the probability of a fish having a length between $l-\frac{1}{2}$ and $l+\frac{1}{2}$. In the case of different length groups, the modification should be obvious.

Now, suppose the true proportion of fish in age group a is π_a. In this case the proportion of fish in length group l, across all ages becomes the proportion in length group equation.

Definition 3.2. Proportion in length group equation:

$$\sum_a \pi_a \left\{ \Phi\left(\frac{(l+\frac{1}{2})-\mu_a}{\sigma_a}\right) - \Phi\left(\frac{(l-\frac{1}{2})-\mu_a}{\sigma_a}\right) \right\}.$$

π_a=proportion of fish in each age group
μ_a=mean length at age
σ_a=standard deviation

Given data (observations) on the proportions at length, those can be compared to the theoretical proportions. A formal statistical approach would be to estimate the unknown parameters by minimizing the discrepancy between the observed and theoretical values. This can be done using any number of discrepancy measures.

Since it is usually difficult to estimate all parameters at once, it is common practice to fix all standard deviations at a fixed value and to fix all mean lengths to some reasonable guesses or values from other sources. Having fixed these, the observer can estimate all the proportions. The next step is to investigate which of the fixed values can be estimated using either these proportions set at the first estimates or estimated with the proportions, starting with the first estimates as initial values.

3.5.2 Examples

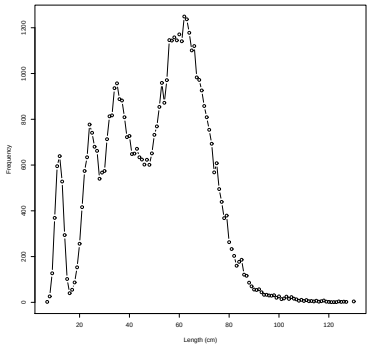
Example 3.2. Suppose the current length group is stored in the variable "lgrp", the proportional number in each age group is "pi", the mean length at age is "mu"and the standard deviation is "sigma". In this case the following R command will compute the proportion in length group "lgrp".

```
sum(pi*(pnorm((lgrp+0.5-mu)/sigma)-pnorm((lgrp-0.5-mu)/sigma)))
```

Notice how every arithmetic operation within this expression is an operation on vectors and the final summation is across all ages. The reader should test and study this command line since it will be used in the following.

3.6 Statistical estimation of proportions at age

Parameters: π_a, μ_a, σ_a
Need some criteria, e.g. $\sum (y_l - \hat{y}_l)^2$
Where y_l is the measured proportion in cell l and $\hat{y}_l$ is the modeled proportion



Example length distribution. Cod in 1990 survey of Icelandic waters.

3.6.1 Details

The proper approach to the issue of age-segregating the catch at length is to assume a statistical model and to separate the length distributions by age according to this model.

The most common model is to assume that the distribution of lengths within an age group is Gaussian and to estimate the parameters of these length distributions along with the proportion in each age group using maximum likelihood or some derivative method thereof.

When the length measurements are in integer units (e.g. 1cm groups) this can easily be done in spreadsheets, usually through discrete approximations to the Gaussian density, but as noted above, a more formal statistical approach using the cumulative distribution function is more appropriate.

The nonlinear minimization becomes quite tricky at times and considerable thought needs to be given into the sequence in which parameters are estimated.

The large number of parameters implies further problems in estimation due to confounding. This can sometimes be alleviated by reducing the number of parameters, e.g. by assuming some standard deviations to be constant or by assuming a growth curve to apply to the mean lengths at age. Such parsimony may lead to considerably more stable results.

In many cases data on mean length at age and on standard deviations at age can be carried across years. Thus, the estimation may be reduced to only estimating the annual proportions in each age group.

In some cases several years worth of data can be combined into a single estimation process. In this case the method is transformed into an assessment procedure, dealt with elsewhere.

3.6.2 Handout

Given that the predicted proportional length distribution is given by

$$\hat{y}_l = \sum_a \pi_a \left\{ \Phi\left(\frac{(l+\frac{1}{2}) - \mu_a}{\sigma_a} \right) - \Phi\left(\frac{(l-\frac{1}{2}) - \mu_a}{\sigma_a} \right) \right\}.$$

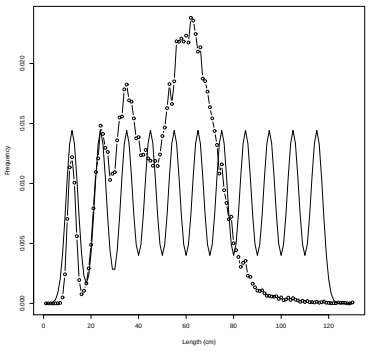
a criterion is needed to estimate the set of unknown parameters, : π_a, μ_a, σ_a.
Some criterion is needed to describe the fit. The simplest of these is a straightforward sum of squares,

$$\sum (y_l - \hat{y}_l)^2 .$$

Where y_l is the measured (observed) proportion in cell l and $\hat{y}_l$ is the modeled proportion, given above.

In order to fit anything to the data one first needs to define the number of age groups to be used. In this particular example three initial modes can be clearly identified. Looking at the actual data it is seen that these occur at 12, 24 and 35cm. Given that these increments amount to 12 and 11 cm and will diminish, it is not too unreasonable to expect that the remaining range of 95cm (=130-35) may correspond to another 8 age groups (or more), so assuming $n_a = 11$ age groups should not be too far-fetched.

3.7 Setting initial values



Example length distribution with initial values for fitted distribution.

3.7.1 Details

Initial values for all parameters are needed before actual estimation can start.

In this example initial proportions are set equal, mean length at age starts out at the peaks for a few of the youngest ages and are then equally spaced. The number of potential age groups is set to 11. The standard deviation is set to 2.5 for all ages, which looks about right for the youngest where it is reasonably clear.

3.7.2 Examples

Example 3.3. For the cod data, initial values for the mean length at age can be set at the first few modes and then with a constant difference, e.g. at $\mu_0 = (12, 24, 35, 45, 55, 65, 75, 85, 95, 105, 115)$.

To initialize the standard deviation of length at age, note that the first mode is at 12cm and the lowest value is at 7cm, with a difference of 5cm. If this range corresponds to 95% probability (bi-directional two standard deviations), then the standard deviation for the first age group should be about 2.5cm. The simple initialization places this as a constant standard deviation for all length groups.

As is seen in the figure, the resulting initial length distribution is very far from the true one. However, it does not need to be very close since the first step will be to modify the proportions.

3.7.3 Handout

Before an attempt is made to minimize the sum, initial values must be set. Some general guidelines can be given, though exceptions will occur.

Proportions at age are easiest initialized at equal values, $\pi_{a,0} = (1/n_a)$.

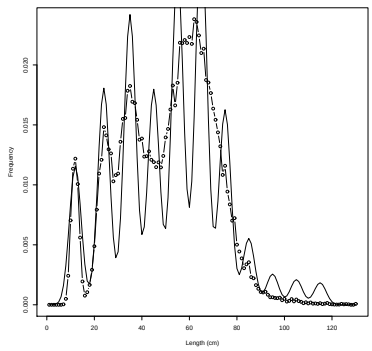
The mean lengths at age for the first few ages can often be determined from the length distribution. Naturally there will be obvious exceptions such as when a year-class is so small as not to be identifiable in the length distribution (a gap will be apparent between the peaks). A different method altogether is to use a von Bertalanffy curve and use starting values of K and L_∞ which pass through the peaks in some manner (see the following sections).

The standard deviation of length at age can often be estimated for the youngest age. If this age group stands out as a peak, then the entire width of this peak can be assumed to correspond to two standard deviations, giving the required estimate.

3.8 Estimating proportions alone

The proportions are easiest to estimate and can be estimated for given values of the mean and standard deviation.

The results from this first estimation part will typically clear up what needs to be done about standard deviations and means.



Example length distribution. Cod in 1990 survey of Icelandic waters.

3.8.1 Examples

Example 3.4. For the cod data an obvious next step is to estimate the proportions first since the initial values are way off. The figure indicates the results from estimating the proportions, given the assumed values for the mean lengths and standard deviations.

Note that although the result is much better than the initial values, there is still a long way to go before this can be called a good fit.

3.8.2 Handout

First steps:
Here one could set up a function to calculate the SSE as a function of the proportions alone, for fixed values of the mean length and standard deviation at age.

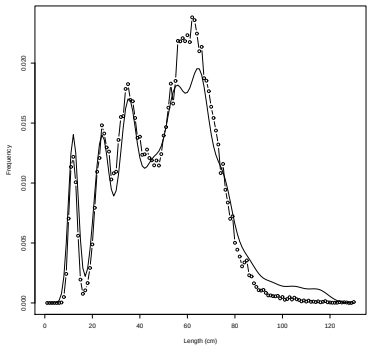
http://tutor-web.net/fish/fish5103growth/lecture30/ssewithfixedlength.r

3.9 Typical parameter reduction - sigma

Commonly assume equal standard deviations

Often the standard deviation for the youngest is clearly lower than for the oldest

Possibly estimate one or a few for the younger ages



Example length distribution. Cod in 1990 survey of Icelandic waters. Sigma values estimated after proportions.

3.9.1 Details

Consider again the typical estimation problem when trying to estimate all proportions, mean lengths and standard deviations from a single length distribution.

If there are n_a age groups, the proportions must sum to one so there will be $n_a - 1$ independently estimated proportions, π_a. There can usually be no other restrictions on these since they correspond to year-class sizes which are usually completely unknown. There must be on mean length at age per age group, giving a further n_a unknown parameters μ_a.

There are also n_a standard deviations, σ_a, so the total number of unknown parameters are $3n_a - 1$, which is usually an uncomfortably large number to be estimated from a single length distribution.

The first obvious problem is how one should estimate all of the standard deviations since these will be very poorly estimated for all but possibly the youngest 1-3 age groups. Thus, if there are clearly distinguishable peaks in the length distribution corresponding to the

youngest age groups it may be possible to estimate the width or standard deviation corresponding to these peaks. It is common practice to try to reduce the number of standard deviations to be estimated to a low number, from one to four depending on the amount of information obviously available in the length distributions.

A second problem, addressed in a later section, is how one can ensure that $p_a \geq 0$ and $0 \leq \mu_1 \leq \mu_2 \leq \ldots$, not to mention that the mean lengths, μ_a, should be "reasonable", i.e. these values should not increase in arbitrary leaps and bounds but must correspond to a realistic growth curve.

3.9.2 Examples

Example 3.5. In the cod example there are a large number of age groups. It follows that it is very unlikely that all the standard deviations can be estimated. On the other hand it is very clear that they are not constant.

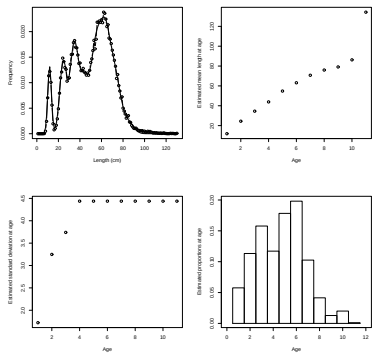
It therefore makes sense to estimate some of these but not all. For example, one can try to estimate the first four standard deviations and assume that standard deviations from age four onward are all the same. Alternative assumptions could of course also be tested.

3.9.3 Handout

The following R code snippet defines a sum-of-squares function for estimation of a fixed few standard deviations.

http://tutor-web.net/fish/fish5103growth/lecture30/sumofsquaresfixedlength.r

3.10 Full run



Example length distribution. Cod in 1990 survey of Icelandic waters. Parameters estimated in twice-repeated sequence.

3.10.1 Details

For the cod example, the previous sections have indicated how it is possible to first obtain initial values, then estimate proportions only, and subsequently the standard deviations, each time fixing all other parameters. Naturally the next step is to estimate the mean lengths at age, using the previously estimated values of proportions and standard deviation.

Given that each of these estimations is only done for a subset of parameters, it is important to repeat the procedure at least once and it is preferable to end the process by estimating the entire set of parameters simultaneously in order to ensure that a "best" estimate is found.

3.10.2 Example

Example 3.6. Coding in R: In addition to the previously defined functions `sseprop` and `ssesigma`, the following function `ssemu` is required to estimate the mean lengths at age, given the values of π_a and σ_a.

```
# Define a function to evaluate the fit of different mu vectors.
ssemu<-function(muvec){
  fit<-rep(0,130)
  for(lgrp in 1:130){
    fit[lgrp]<-sum(pi*(pnorm((lgrp+0.5-muvec)/sigma)-pnorm((lgrp-0.5-
        muvec)/sigma)))
  }
  sse<-1e6*sum((dat-fit)^2)
  plot(lgrps,dat,type='b',lwd=2)
  lines(lgrps,fit,type='l',lwd=2,col="red")
  cat("SSE=",sse,"\n")
  return(sse)
}
```

Suppose this function is also defined in the file `functions.r`, along with the previous `sse`-functions.

The same `init` file as before can be used but a longer set of function calls need to be used since now the purpose is to repeatedly call each function.

3.11 Parameter reduction - using a growth curve

May want to use a von B growth curve in place of μ_a

3.11.1 Details

In applications several estimation issues may arise. In particular it may not be feasible to estimate individual length at age for all ages, even though it may be possible to clearly discern a few peaks for the youngest ages. In this case one may want to use a von Bertalanffy growth curve, i.e. assume

$$\mu_a = L_\infty \left(1 - e^{-K(a-t_0)}\right)$$

so the mean length at age is based on only 3 parameters rather than tring to freely estimate a mean length for every age.

3.11.2 Example

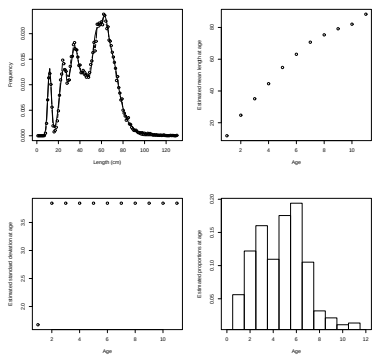
Example 3.7. R coding: A typical function for estimating the vonB growth curve could be the following. As in the previous sections, this particular function is geared towards the cod example, with a fixed 11 ages and length groups from 1 to 130 cm.

http://tutor-web.net/fish/fish5103growth/lecture30/vonbgrowthcurve.r

3.12 Caveats

Uncertainty estimation
Time of sampling - fixed-point (survey) or continuous (catches)
Time of spawning - short interval or continuous or biomodal
etc

3.13 The next steps



3.13.1 Details

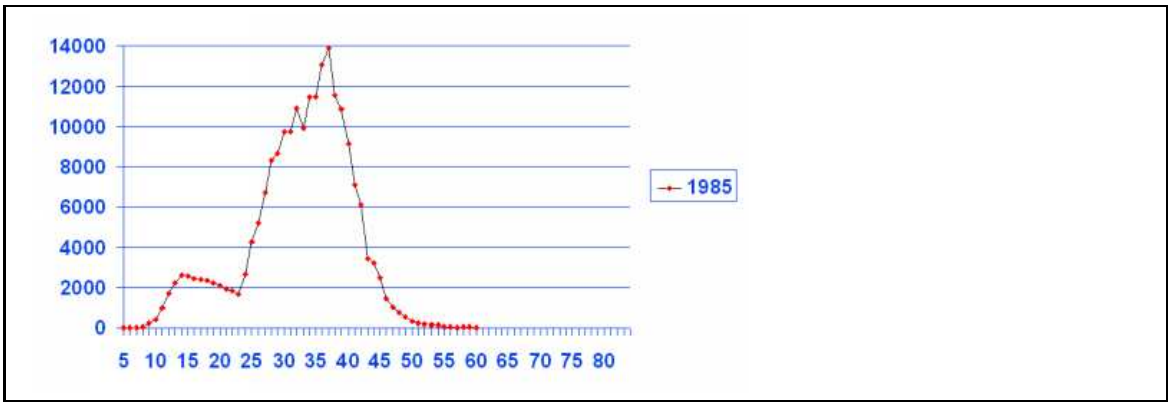
The next steps are to make the previous routines generic and to set up an `R` library. The routines should, for example, not contain exactly 11 age groups and 130 length classes etc. Routines should also be added to provide reasonable initial values for all parameters or at least initialize from a minimal set of assumptions.

3.13.2 Assignment

Assignment 3.1. Modify the routines in this section and and use them on another stock.
Collection of generic R code
`http://tutor-web.net/fish/fish5103growth/lecture30/base30.dat`
`http://tutor-web.net/fish/fish5103growth/lecture30/init1.r`
`http://tutor-web.net/fish/fish5103growth/lecture30/functions1.r`

4 Case studies in analysis of length data

4.1 Two yearclasses or many: A case study, part 1



4.1.1 Details

Some care must be exercised when attempting to read ages from length distributions, as the following case study will show.

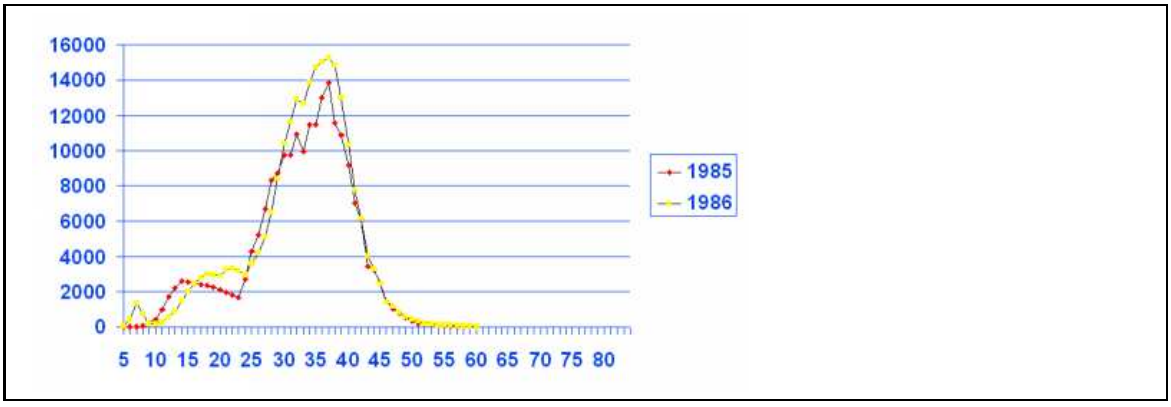
Before anything else is known about a species, it is typically sampled for length measurements, providing a length distribution such as that above.

This particular length distribution is from a species in Icelandic waters. Assume for the moment that these data are the only data available on this species. Such a situation is not all that uncommon especially when research is initiated on a new stock.

Before anything else is known about the species, there is no reason to assume anything about the age structure. In particular, the length distribution in this case gives no indication as to whether the two identifiable peaks correspond to two cohorts, and even if they are, then it is not clear whether the two cohorts are adjacent in time or far apart.

In tropical waters such two peaks might even illustrate a single year-class spawned during two different spawning seasons within a year.

4.2 Two yearclasses or many: A case study, part 2

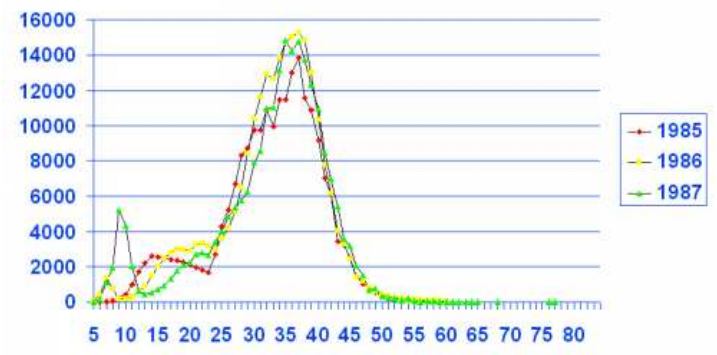


4.2.1 Details

When data for two adjacent years are considered, it is seen that the pattern of two major peaks in the length distribution is repeated. However, it is also seen that a new peak at a very small length (7 cm) appears.

The question is now raised, whether there are three year-classes of a very fast growing species or whether some other interpretation exists.

4.3 Two yearclasses or many: A case study concluded

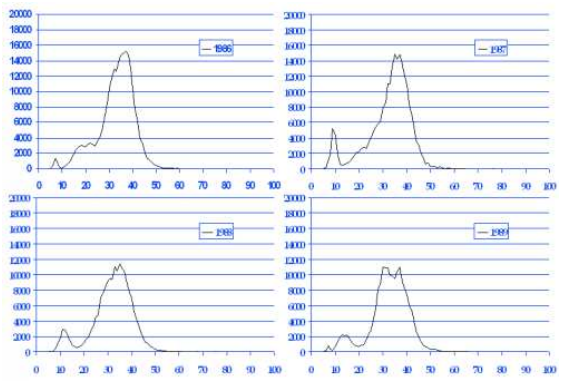


4.3.1 Details

As more years are added to the picture, the required detail emerges. With three years of data it is now clear that the peaks are indeed cohorts or groups of cohorts. This is clear since the peaks move to the right by about 2 cm per year.

Thus, it is concluded that this is not a fast-growing species but a very slow growing one. Further, the long distance between the peaks is an indication of tremendous year-class variation.

4.4 Two yearclasses or many: The actual situation



4.4.1 Details

The species in question is redfish. This is a very slow growing species with enormous year-class variation.

Early attempts at age determination indicated that there was very little year-class variation. A simple analysis of length distributions, such as this one, clearly invalidates those age readings.

Experienced fishermen and scientists had problems in figuring out details in the dynamics of this stock. For example, a catch of juveniles from this stock appears as a whole bunch of small fish, around 10cm. It is not until the data have been collected and the length distributions drawn up as a time series, that the growth pattern emerges and indicates that the pile of small fish on deck is not always 10cm but grows by about 2cm per year.

5 Length-weight relationships

5.1 Estimating the relationship between length and weight

5.1.1 Details

The simple approach to estimating the relationship between length and weight is to log-transform lengths and weights, followed by a simple linear regression. However, the log-transformation causes a bias which needs to be investigated.

Alternative methods include using generalized linear models with e.g. a gamma distribution and a log link.

It in known that there are often problems with how a single relationship fits the small fish or the large fish. Different relationships may be needed for different age ranges, or smoothing functions may be needed to avoid consistently under- or over-predicting the weight of the extreme size classes.

The models can be formally tested using a lack-of-fit test for the mean function, tests of distributional assumptions, tests for outliers and so forth.

The entire battery of test mechanisms for simple linear regression models is available for the length-weight relationships.

It should be noted that the primary hypothesis is not whether the slope in the log-log regression is zero, but rather is the slope 3. This hypothesis is usually rejected and that implies that the usual definition of a conditional factor (as w/l^3) is not appropriate.

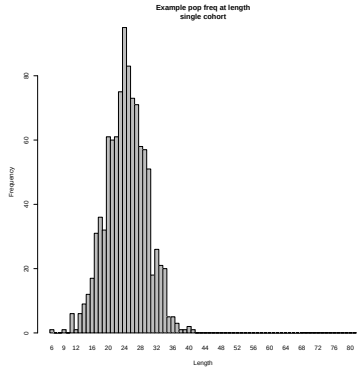
Finally, the R^2-value, reported by popular spreadsheets when a “power-curve” is fitted, has no meaning for the curve and should not be used. The spreadsheet user should be aware of this and of the fact that the resulting curve may be seriously biased.

5.1.2 Examples

Example 5.1. http://tutor-web.net/fish/fish5103growth/lecture50/lwrelationship.r

6 Modelling the development of a length distribution

6.1 A length distribution



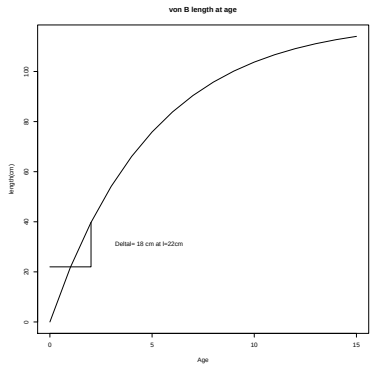
A simulated length distribution.

6.1.1 Details

A (discrete) model of length distribution development from growth needs to specify how each length group progresses from one group to the next via growth.

The following will ignore mortality and other issues, in order to focus on the effect of growth alone.

6.2 A growth curve

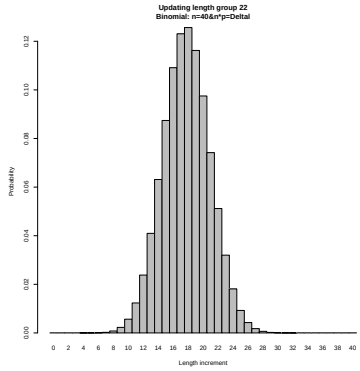


An example of a von Bertalanffy growth curve.

6.2.1 Details

If a general growth curve can be assumed, then that can be used to determine the mean growth for a group of fish. Specifically, a von Bertalanffy growth curve can determine the mean growth of each length group.

6.3 The updating distribution



The updating distribution describes the frequency of movements from a fixed length cell into other cells.

6.3.1 Details

Given the number of fish in a given length group and a mean growth, an updating distribution should be used to describe how the fish move into adjacent length groups so that mean growth can be determined.

The simplest model would simply move all the fish by the same number of length cells. The immediate problem with this is that typically the growth is by a fraction, not an integer.

This could be fixed by moving the fish into the length group immediately above or below by the proportions dictated by the mean growth. This was, for example, done in MULTSPEC.

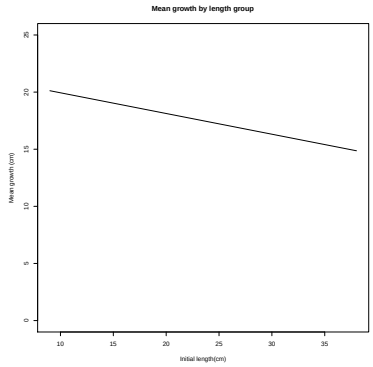
However, some information exists about how the distribution of fish at length changes in time, e.g. for a given cohort. Similarly, tagging data gives direct information on how the growth corresponds to a spread across several length groups.

It is therefore useful to consider growth in terms of an updating distribution.

6.3.2 Examples

Example 6.1. Take a single length group and model how this should grow into adjacent cells, restricted in such a way as to provide the proper mean growth. The example shows how a binomial distribution can be used.

6.4 Growth from length at age

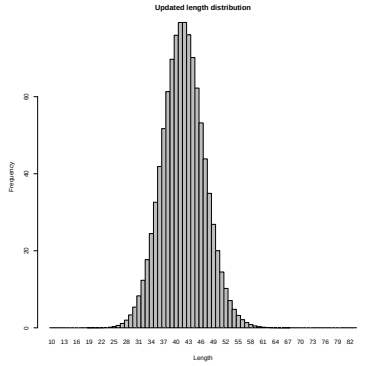


Varying initial lengths will correspond to different growth under the von Bertalanffy growth curve.

6.4.1 Details

The growth function (or length-at-age function, rather) will predicate different growth patterns for different initial lengths. Thus the updating distribution will be different for different initial length groups.

6.5 An updated length distribution

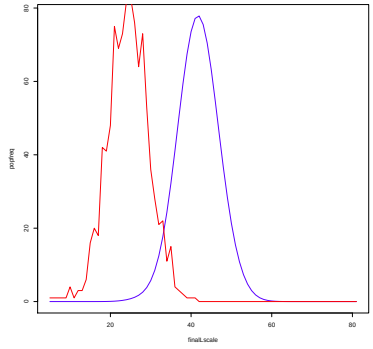


The simulated data transforms into a smoother length distribution since each target length group contains a sum of parts from several other length groups.

6.5.1 Details

When the updating mechanism has been used to move each length group, the result is a new, shifted, length distribution.

6.6 The update as a shifting smoother

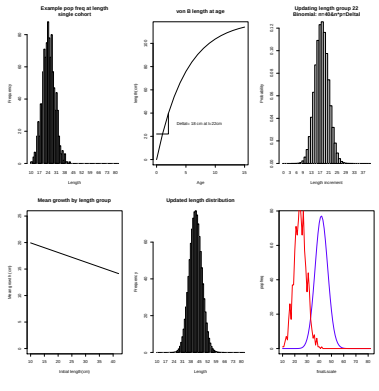


Notice how the updating mechanism shifted and smoothed the initial length distribution.

6.6.1 Details

An updating procedure will split each initial length into several adjacent ones, each corresponding to some growth. This corresponds to a combination of shifting, smoothing, and spreading out of the length distribution.

6.7 An example of an updating model



Caricature of an update mechanism

6.7.1 Details

A length updating mechanism is implemented by specifying a method for how much fish of a specified length should grow and then prescribing how a mean growth should correspond to a distribution of fish onto different length groups.

6.7.2 Example

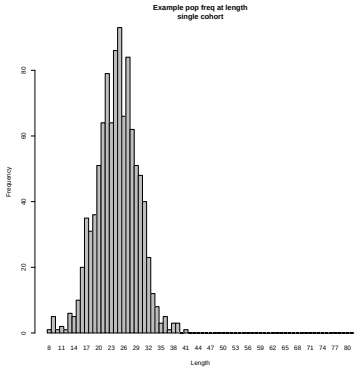
Example 6.2. The entire updating mechanism can be demonstrated using the following R code:
http://tutor-web.net/fish/fish5103growth/lecture60/updating-mechanism-in-gadget.r

7 Using length data in population models

7.1 Introduction

Several approaches exist:

- Cohort slicing and then VPA-style
- Just use lengths for recruits and then statistical models
- Model length dist as a part of pdy model



A simulated length distribution.

7.1.1 Details

Several approaches exist to include length data in population models.

- Cohort slicing and then VPA-style: This is a very simple method and can always be used if there is basic information about the growth of the species, i.e. some information on where the length distributions should be sliced. Naturally, more elaborate methods than cohort slicing could also be used.
- Use lengths of recruits and then statistical models: For very many species the first one or two age groups are clearly separated from the rest, but the older ages are all in one lump. General population dynamics models can easily be fitted to data when there is an index of recruitment and a separate index of the adult population.
- Model the length distribution as a part of a population dynamics model: In this case an internal population dynamics model is used to predict the correct length distribution. The measured length distribution is then used for comparison. Parameters of the population dynamics model are estimated by ensuring that the predicted length distribution is close to the measured one.

7.1.2 Examples

Example 7.1. Suppose we want to track the length distribution of a single cohort (for simplicity).

7.1.3 Assignment

Assignment 7.1. When estimating parameters using a non–linear minimizer it is preferable to provide 'reasonable' starting values. A way to do this for the von Bertalanffy equations is as follows:

$$L_t = L_\infty(1 - exp(-K[t - t_0]))$$

Assume $t_0 = 0$, then,

$$L_t = L_\infty(1 - exp(-Kt)) \tag{1}$$

$$L_t = L_\infty - L_\infty exp(-Kt) \tag{2}$$

$$L_\infty - L_t = L_\infty exp(-Kt) \tag{3}$$

Substitute L_{t+1} for L_t in equation 2.

$$L_{t+1} - L_t = L_\infty exp(-K(t+1)) - L_\infty exp(-Kt) \tag{4}$$

Which can be written as:

$$L_{t+1} - L_t = -L_\infty exp(-Kt)(1 - exp(-K)) \tag{5}$$

Substitute equation 4 into equation 5.

$$L_{t+1} - L_t = (L_\infty - L_t)(1 - exp(-K))$$

L_{t+1} can be written as a function of L_t.

$$L_{t+1} = L_\infty(1 - exp(-K)) + L_t exp(-K) \quad (6)$$

L_{t+1} can be plotted against L_t and a linear regression model fitted.

$$L_{t+1} = a + bL_t$$

where $a = L_\infty(1 - exp(-K))$ and $b = exp(-K)$. K and L_∞ can then be estimated from:

$$K = -\ln(b)$$
$$L_\infty = a/(1-b)$$

von Bertalanffy — predicting
R can be used to calculate the predicted length at age (`lhat`) from the von Bertalanffy equation:

```
lhat <- Linf*(1-exp(-K*(a-t0)))
```

where `Linf` , `K` and `t0` are constants and `a` is the age(and can be a vector).

The equation can also be stored as a function:

```
vonb <- function(b) {
  Linf <- b[1]
  K <- b[2]
  t0 <- b[3]
  lhat <- Linf*(1-exp(-K*(a-t0)))
  return(lhat)
}
```

where `b` is a vector containing L_∞, K, t_0. The required ages a are in another predefined vector.

- Write a function in R to predict length at age with $L_\infty = 160$, $K = 0.09$ and $t_0 = 0$ for fish aged 2 to 14.
- Plot the predicted lengths at age.
- Calculate the growth curves for different parameters and compare on the same plot.

von Bertalanffy — estimating
Given data on mean length at age, the von Bertalanffy parameters can be estimated by finding the parameters which minimize the difference between the observed mean lengths at age and those predicted from the von Bertalanffy equation.

To calculate the sum of squared errors for a set of parameters use a function like this:

```
vb.sse <- function(b){
  lhat <- vonb(b)
  sse <- sum((l-lhat)^2)
  return(sse)
}
```

This function combines the function to predict the von Bertalanffy growth curve (`vonb`) with the calculation of the difference between the observed growth curve and predicted growth curve `sse`.

The best fit to the data is from the parameters giving the lowest sum of squared errors. To estimate these parameters there is an R function `nlm` which is used like this:

```
est <- nlm(vb.sse, c(100, 0.1, 0))
```

It is better to use:

```
est <- nlm(vb.sse, c(100, 0.1, 0), typsize = c(100, 0.1,0.001))
```

where `typsize` is an estimate of the order of magnitude of the optimized parameters. Using this should reduce the number of iterations required, increase the chance an optimum will be found and make the optimization more robust. There are many other options in `nlm` which are explained in help.

The output of `nlm` includes

```
est$estimate # the parameter estimates
est$minimum # the minimum sse
est$code # a code which tells you the status
                # (i.e. whether an optimum has been found)
```

To see the search being done by `nlm` add

```
lines(a, lhat, col=2)
```

to `vb.sse` and set up a plot before running `nlm` e.g.

```
plot(a, vonb(c(Linf,K,to)), type="n")
```

von Bertalanffy – fitting to data

- Given these data:

  ```
  L <- c(18, 24, 29, 32, 35)
  t <- c(1.5, 2.5, 3.5, 4.5, 5.5)
  ```

  - Estimate the von Bertalanffy parameters
  - Plot L_t against L_{t+1}
  - Fit a linear regression line using `lm`
  - Use the estimates of K and L_∞ from the linear regression to start the non-linear minimization.

- A dataset including haddock age and length is available from:

  http://tutor-web.net/fish/fish5103growth/haddockage.dat

  - Using these data, calculate mean length at age (`tapply` and fit a von Bertalanffy growth curve assuming age 1 is equivalent to $t = 1$ *etc*.
  - Add the fitted growth curve to a plot of the observed data
  - How many fish are there at each age? Does eliminating from the analysis ages with a very small number of fish affect the estimated growth curve?

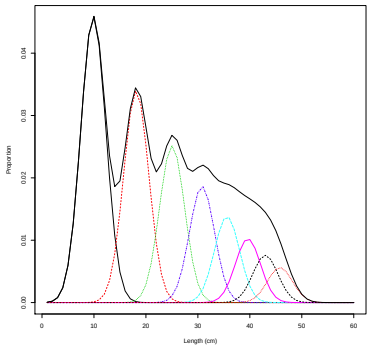
7.2 Simulating a length distribution

Population numbers can be simulated using the stock equation

Mean length at age can be generated from a von Bertalanffy curve

The distribution of length at age can be generated from the mean and standard deviation assuming a Gaussian distribution

The population length distribution is generated by adding the age-based length distributions



Simulated proportional length distribution of individual age groups and the total population in equilibrium.

7.2.1 Details

It is fairly simple to simulate population length distributions based on assumptions such as a fixed year-class size and growth according to a von Bertalanffy curve.

These simulated curves can be used in an estimation procedure to evaluate whether the same growth can be extracted. If one generates a length distribution which has the same overall flavor as a particular sample, this type of fitting to simulated data provides very simple tests of whether there is any chance of extracting information from the observer length distribution.

Noise can then be added to evaluate uncertainty in the estimation procedure.

Most length distributions are aggregates of several samples. A better method to estimate uncertainty would be to bootstrap the entire length distributions and obtain repeated estimates of growth and proportions in the age groups.

Note that it is not sufficient to bootstrap (resample) on individual fish since the fish in a given length sample are correlated and therefore entire samples needs to be bootstrapped.

7.2.2 Examples

Example 7.2. The plot gives the proportional length distribution for individual year-classes and the total population, when the stock equation is used to generate year-class size from a fixed number of recruits and the von Bertalanffy growth curve is used to generate mean length at age.

The simulated figure is typical of what an observed length-frequency plot can look like.

7.2.3 Assignment

Assignment 7.2. Models of length distribution
A single age

- If the length distribution of an age-class is Gaussian (normal) with mean length $\mu = 23$ and $\sigma = 1.4$ then it could look like this:

```
# generating a point value for every cm
le <- seq(15,30,1)
plot(le, dnorm(le, 23, 1.4), lwd=2)
lines(le, dnorm(le, 23, 1.4))
```

- If a length group is assumed to go from -0.5 to +0.5 then it is more appropriate to sum (integrate) the distribution within this region rather than use the density of the mid–point. The *cumulative distribution* is then used:

```
# the cumulative distribution is a continual sum of the density
    distribution
le <- seq(15,30,1)
plot(le, pnorm((le-23)/1.4), type="o")
# the density distn from the cumulative
leA <- seq(15.5,30.5,1) # upper point
leB <- seq(14.5,29.5,1) # lower point
# to check leA and leB are correct
rbind(leA,leB)
d2 <- pnorm((leA-23)/1.4) - pnorm((leB-23)/1.4)
# the values from the cumulative distn
plot(le, d2, type="o")
# the values from the density distn
points(le, dnorm(le, 23, 1.4), col=2)
```

Several ages
To model several ages, the mean length at each age can be defined in a vector eg `mu <- c(23, 34, 42, 50)` or a growth model can be used to define the mean length at age.

- A population containing several length distributions can be simulated like this:
- The mean length at age is predicted using the von Bertalanffy model.
- The standard deviation of length at age is fixed.
- The number at age is predicted using the stock equation.

```
# generating the mean length from von Bertalanffy
a <- 1:8 # 8 age groups
mu <- vonb(c(60,0.18,0))
sdev <- rep(2.4,length(a))
le <- 1:60
leA <- seq(1.5,60.5,1) # upper point
leB <- seq(0.5,59.5,1) # lower point

# the proportion at age assuming no fishing and M = 0.3
N <- 1
M <- 0.3
for(i in a){
  N <- c(N, exp(-M)*N[i])
```

```
}

Ntot <- sum(N)
p <- N/Ntot

# use a loop to calculate length distribution for each age
# and store in rows

# the length distribution * proportion at age

ldist <- NULL
for(i in a) {
  d2 <- (pnorm((leA-mu[i])/sdev[i]) - pnorm((leB-mu[i])/sdev[i]))*p[
     i]
  ldist <- rbind(ldist, d2)
}

# these can be plotted individually
matplot(le, t(ldist), type="l")
# and the length distribution would look like
plot(le, apply(ldist, 2, sum), type="l")
```

Practicals

- Plot the length distributions for different von Bertalanffy growth curves and with different patterns of mortality at age (eg with only natural mortality on younger fish but fishing mortality on older fish.

8 Using length data in population models

8.1 Introduction

Several approaches exist:

- Cohort slicing and then VPA-style
- Just use lengths for recruits and then statistical models
- Model length distribution as a part of population dynmaics model

8.1.1 Details

Several approaches exist:

- Cohort slicing or other age-segregation and then VPA-style assessment
- Use lengths for recruits, or R and B, and then statistical models
- Model length distribution as a part of population dynamics model

This topic is enough for a complete course in fisheries. Importantly, if annual length distributions exist, then there is potentially some information about age groups. Many assessment procedures are available, which typically have an internal age and growth structure, which predicts annual length distributions, e.g. Gadget.