

## Analyses of length and age distributions using continuation-ratio logits

**Rindorf, Anna; Lewy, Peter**

*Published in:*  
Canadian Journal of Fisheries and Aquatic Sciences

*Publication date:*  
2001

*Document Version*  
Peer reviewed version

[Link back to DTU Orbit](#)

*Citation (APA):*  
Rindorf, A., & Lewy, P. (2001). Analyses of length and age distributions using continuation-ratio logits. Canadian Journal of Fisheries and Aquatic Sciences, 58(6), 1141-1152.

## DTU Library

Technical Information Center of Denmark

---

### General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

## **Analyses of length and age distributions using continuation-ratio logits**

Anna Rindorf\* and Peter Lewy

\* Author for correspondence: University of Copenhagen, c/o Danish Institute for Fisheries Research, Charlottenlund Castle, DK2920 Charlottenlund, Denmark. E-mail: [ar@dfu.min.dk](mailto:ar@dfu.min.dk)

## **Abstract**

Sampling of length and age distributions of catches is important for the assessment of commercially fished stocks. This paper presents a new method for statistical analyses and comparisons of length and age distributions based on generalised linear models of continuation-ratio logits. The method allows statistical testing of the effects of both continuous and discrete variables. Further, by utilising the smoothness of length and age distributions as a function of length, the method provides more accurate estimates of these distributions than traditional methods. The observations are assumed to be multinomial distributed, but cases in which the variance exceeds that of this distribution may also be analysed. The implementation of the method in existing statistical analysis software is straightforward and is demonstrated using length and age distributions of lesser sandeel, *Ammodytes marinus* Raitt.

## Introduction

Length and age compositions of numerous commercially fished stocks are analysed on a regular basis to assess recruitment, stock biomass and other aspects of the state of the stock. Samples for this purpose are obtained either by commercial catch sampling, research surveys or both. In a typical sampling programme, catch of a species is sampled by recording the total number or weight caught of the species, the length distribution of a sub-sample of the catch and the age at length of a sub-sample of the length sample (Cotter 1998).

The multinomial distribution is frequently used to describe length and age distributions and most authors compare these distributions by aid of the  $\chi^2$ -test of homogeneity (Baird 1983; Zwanenburg and Smith 1983; Engås and Soldal 1992). However, the  $\chi^2$ -distribution has the disadvantage of being an inaccurate approximation when the expected number of outcomes in a category is less than five (Cramér 1946) and the variance in the collected data is often greater than can be described by the multinomial distribution (Smith and Maguire 1983, Williams and Quinn 1998).

The calculation of catch in numbers at age from numbers at length is usually based on an age-length-key. The traditional key is based on a two-way table of age and length in which the entries are number of fish at length and age (Fridriksson 1934). This key does not take the knowledge of fish growth into account. However, this relationship has been included in some investigations by assuming the length distribution of a particular age group to be normal (Schnute and Fournier 1980; Labonté 1983; Gudmundsdóttir et al. 1988).

This paper presents a new method to statistically analyse and compare samples of age and length distributions from different geographical areas, time periods, laboratories, etc. The object of the analyses is to enable estimation of age and length distributions with the

lowest possible uncertainty. One important step on the way is the decision on whether or not to pool distributions from different strata. This multinomially based method provides a statistical tool for this decision comparable to the Anova in the case of normal distributed data.

The method is based on continuation-ratio logits as presented by Agresti (1990) and previously applied by Kvist et al. (2000) to analyse age distributions. However, the method presented here has several additional advantages to previous applications. It can take the smoothness of length distributions into account even in cases where the length distribution is skewed or polymodal, and can be used in cases where the variance of the observations exceeds that of the multinomial distribution. The method utilises the knowledge of increasing mean fish age as a function of size without any assumptions on growth apart from that mean length of older fish is greater than that of younger. Further, it can easily be implemented in existing statistical analysis software, and the effect of both categorical and continuous variables on the distribution may be tested. Methods to obtain the accuracy of estimated length or age distributions are provided as well. The method is demonstrated using survey data on length and age composition for lesser sandeel, *Ammodytes marinus* Raitt.

## **Analysing ordinal multinomially distributed variables by continuation-ratio logits**

Generalised linear models are a convenient tool for statistical analysis of variables following distributions in the exponential family (McCullough and Nelder 1989). Applying these to multinomial data enable model selection, estimation and testing techniques developed for generalised linear models to be employed. It is thus possible to simultaneously analyse several different length or age distributions. This implies that the

small sample problems can be reduced. Further, the ordinal nature of length and age measurements makes the application of the continuation-ratio logits as presented by Agresti (1990) possible. This has several advantages when analysing length and age distributions as shall be shown in the following.

Continuation-ratio logits can be used for any multinomial variable for which the outcomes are ordinally scaled. Assume that we have observations in ordinally scaled groups  $i$  and let  $p_i$  be the probability of the outcome falling in group  $i$ . The continuation-ratio logits are then defined as (Agresti 1990):

$$\log\left(\frac{p_i}{p_{i+1} + \dots + p_n}\right), \quad i = 1, \dots, n-1$$

This logit corresponds to the binomial logit

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right)$$

where

$$\pi_i = \frac{p_i}{p_i + \dots + p_n} = P(\text{index} = i \mid \text{index} \geq i)$$

for  $i=1, \dots, n-1$ . That is,  $\pi_i$  is the probability of the outcome being of index  $i$  conditioned on the outcome being of index  $i$  or greater.

The continuation-ratio logits can be modelled separately or simultaneously in a generalised linear model of a binomially distributed variable, as the likelihood of a multinomial trial can be expressed as  $\prod_{i=1}^{l=k-1} b[\pi_i, n_i, n_{i+}]$  where  $n_i$  is the number of outcomes in category  $i$ ,  $n_{i+}$  is the number of outcomes in category  $i$  or greater,  $\pi_i$  is the conditional probability described above, and  $k$  is the number of outcome categories (Agresti 1990). Maximising the multinomial likelihood is thus equal to maximising the product of likelihoods of the conditional binomial distributions of each index. As shown by Agresti (1990), this implies that the continuation-ratio logit of each index may be modelled separately.

The continuation-ratio logits of the different outcomes are not independent observations. However, due to the relationship between the multinomial likelihood and the likelihood of the conditional binomial distributions, maximising the multinomial likelihood is equivalent to maximising the product of the conditional binomial trials simultaneously. Thus, any software implementing maximum likelihood techniques to analyse binomial data may be used to analyse ordinal multinomial data. This includes software designed for analyses of generalised linear models such as SAS<sup>®</sup> and S-Plus<sup>®</sup> (McCullagh and Nelder 1989; SAS Institute Inc. 1996; Mathsoft Inc. 1997). However, the original multinomial data must first be transformed into a data-set containing the conditional binomial observations and the index of the outcome. For instance, if the multinomial data are as seen in the first two columns of Table 1, the data-set to be

analysed in the binomial model contains the index  $i$ , the number of outcomes in this category,  $n_i$  and the number of outcomes of this order or greater,  $n_{i+}$  (columns 1, 2 and 4).

If the data show greater variance than can be accounted for by the binomial variance, this may be incorporated in the generalised linear model by estimating a dispersion parameter (McCullaugh and Nelder 1989). The only constraint imposed by this method is that the dispersion parameter must be the same for all the observations included in the model and thus for all conditional distributions when multinomial data are analysed as described above.

Including the conditional probabilities of all indices simultaneously in one generalised linear model, it is possible to use the index as a continuous explanatory variable. The value of the conditional logit of group  $i$  can then be expressed by

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \lambda(i)$$

where  $\lambda(i)$  is a smooth function of  $i$ . The function  $\lambda(i)$  can e.g. be modelled by a local non-parametric smoother or a polynomial in  $i$ . Alternatively, one model may be estimated for each ordinal group separately, if one does not wish to include the smooth relationship between the conditional probability and the index of the group. The effect of categorical factors on the conditional probability of group  $i$  may also be estimated by generalised linear models.



The model provides estimates of the conditional probabilities,  $\hat{\pi}_i$ , as well as the variances of these estimates. The estimated unconditional probabilities,  $\hat{p}_i$ , can be calculated from the conditional probabilities by applying the formula:

$$\begin{aligned} \hat{p}_1 &= \hat{\pi}_1 \\ \hat{p}_i &= \hat{\pi}_i * \left( 1 - \sum_{j=1}^{i-1} \hat{p}_j \right) = \hat{\pi}_i * \prod_{j=1}^{i-1} (1 - \hat{\pi}_j) \quad , \quad i > 1 \end{aligned}$$

Approximations of the variance of  $\hat{p}_i$  are easily obtained if the variance of  $\hat{\pi}_i$  has been estimated and if the conditional probabilities,  $\hat{\pi}_i$ , are modelled independently of each other. In this case the estimates are uncorrelated and the variance of these can be approximated by the first order Taylor approximation:

$$(1) \quad \begin{aligned} V(\hat{p}_1) &= V(\hat{\pi}_1) \\ V(\hat{p}_i) &\approx \hat{\pi}_i^2 \hat{A}_i^2 \sum_{j=1}^{i-1} \frac{V(\hat{\pi}_j)}{(1 - \hat{\pi}_j)^2} + \hat{A}_i^2 V(\hat{\pi}_i), \quad i > 1 \end{aligned}$$

where

$$\hat{A}_i = \prod_{k=1}^{i-1} (1 - \hat{\pi}_k)$$

Correspondingly, the covariance between the unconditional probabilities,  $COV(\hat{p}_i, \hat{p}_j)$ , can be calculated.

If the conditional probabilities of all groups are analysed simultaneously and one or more parameters are common to all groups, the predicted conditional probabilities,  $\hat{\pi}_i$ , are correlated. This complicates derivation of analytical approximations to the variance of  $\hat{p}_i$ .

In this case simulation studies provide a convenient method for estimating the variance.

A simulation procedure which can be used for this purpose is the parametric bootstrap (Davison and Hinkley 1997). In this type of analysis, observations are simulated from the parameter estimates obtained from the fitted model. The simulated observations are then used to re-estimate the parameters. This is done by repeating the entire procedure of model fitting on a number of simulated data sets. The parameter estimates obtained provide estimates of the simulated variance and mean of these parameters.

### **Estimation of the probabilities and their variance in the multinomial distribution**

The estimated probabilities obtained by the traditional multinomial approach,  $\tilde{p}_i$ , are calculated as

$$\tilde{p}_i = \frac{n_i}{\sum_{j=1}^{j=k} n_j}$$

where  $n_i$  is the number of outcomes in category  $i$  and  $k$  is the number of outcome categories. This probability does thus not take the overdispersion potentially present in the data into account. The variance of the estimate is calculated as

$$(2) \quad V\left(\tilde{p}_i\right) = \frac{\tilde{p}_i(1-\tilde{p}_i)}{\sum_{j=1}^{j=k} n_j}$$

These estimates are referred to as the multinomial estimates in the following.

## **Application of the method**

### **Data analysed**

The data used for demonstration of the method consist of catches of lesser sandeel from three surveys conducted in the Firth of Forth east of Scotland in 1997 and 1998 (Fig. 1). Positions were initially chosen according to prior knowledge of sandeel distribution in the area and the same positions were dredged in both years. A modified scallop dredge was used each year to catch the sandeel while they were buried in the sediment (Winslade 1974). In 1998, the third survey was conducted using a van Ven grab to collect additional sandeel for ageing. This survey was performed at positions 2, 5, 7 and 8 within a week of the dredge survey.

At each station, an average of five dredge hauls of 15 min. durations were carried out in both years. The total content of sandeel of all hauls were weighed separately. If time

permitted, the entire content was counted and length measured to nearest semicentimetre below. Otherwise, a random subsample was counted, measured and weighed.

In 1997, a random subsample of ten sandeel was taken from each semicentimetre group of the dredge catch at each position (provided that the catch exceeded ten sandeel), and the otoliths of these fish were age determined at the Danish Institute of Fisheries Research. In 1998, only a small number of sandeel from the scallop dredge catches were age determined by the Danish Institute. To supplement these samples, sandeels collected in the grab survey were also used in the calculation of the age-length relationship. These samples were age determined by the Institute of Fisheries Research Services, Marine Laboratory, UK, Scotland, and kindly supplied by S. Greenstreet.

#### **Analysing length distribution of catch using continuation-ratio logits**

The observations analysed in this model were the observed length distributions measured in 1 cm groups for each of the hauls taken. Thus, there were five length distributions at each position on average. The number of fish in each length group were assumed to be multinomially distributed. The object of the model developed was to examine whether the length distribution of the catch varied from year to year and/or from position to position. As these length categories are ordinal, the conditional probabilities can be modelled as described above.

Initially, the original multinomial data set was transformed into a data set containing the index of the outcome (length in this case), the number of outcomes in the category in trial  $j$ , the number of outcomes of this index or greater in trial  $j$  and the additional variables year and position. This is equivalent to the transformation illustrated in Table 1. If no fish of index  $i$  or greater were recorded in trial  $j$ , the conditional probabilities of index  $i$  and greater was missing in trial. The length dependent continuation-ratio logits were analysed in a generalised linear model of this binomial data set.

The minimum length observed was 7 cm, and the maximum length for which conditional probabilities were calculated was 19 cm. The model of the continuation-ratio logits included a 7<sup>th</sup> degree polynomial in length and categorical parameters describing effects of year and position:

$$(3) \quad \log\left(\frac{\pi_{l,y,pos}}{1-\pi_{l,y,pos}}\right) = \lambda_{y,pos}(l) = \sum_{n=0}^7 b_n(y, pos)l^n$$

where

$$b_n(y, pos) = b_n + b_{n,y} + b_{n,pos} + b_{n,y,pos}$$

and  $l$  indicates length in group  $l$ ,  $y$  indicates year,  $pos$  indicates position and  $b$  indicate parameters to be estimated in the model. The 7<sup>th</sup> degree polynomial was chosen as the highest order polynomial for which no convergence problems arose. These problems are due to lack of contrast in the data, which makes it difficult or impossible to find a global minimum of the deviance. The model was fitted using the SAS<sup>®</sup> GENMOD-procedure (SAS<sup>®</sup> version 6.12 for Windows<sup>®</sup>, SAS<sup>®</sup> Institute 1996). The dispersion parameter was estimated by the Pearson  $X^2$  statistic divided by the degrees of freedom. This parameter measures the deviation between the variance between samples (hauls) measured and the variance between samples expected in the multinomial distribution. A value of one

indicates that the between haul variation is completely described by random variation within five samples from a multinomial distribution. The model was reduced by eliminating insignificant factors from the model (F-test, 5% level).  $r^2$  values were calculated as the deviance explained by a particular factor divided by the total deviance.

The length proportions,  $\hat{p}_{l,y,pos}$ , estimated from the reduced generalised linear model (including the significant effects of year and position) were used to perform a parametric bootstrap analysis. The number of replicates was 1000, that is, 1000 independent multinomially distributed length composition data sets for all hauls were generated by simulation, using the estimated length distribution at the given year and position as well as the total number of fish length measured in the particular haul as input. For each of these 1000 equivalent data sets, a generalised linear model of the conditional probabilities was used to estimate the length proportions,  $\hat{\rho}_{sim,l,y,pos}$ . The dispersion parameter in these models was fixed at one (no overdispersion). From the 1000 replicates, the mean,  $\hat{\rho}_{l,y,pos}$  and the variance of the estimates were calculated.

The original data were slightly overdispersed as data showed greater variance around the model than could be accounted for by the multinomial variance. The estimation of the length distributions from the original data included this problem by introducing a dispersion parameter (McCullagh and Nelder 1989). However, this overdispersion was not taken into account when simulating data, or when building generalised linear models of the simulated data. Therefore the resulting variances estimated from the simulated data may be somewhat lower than the variances in the original data. However, the comparison between the variance calculated from the simulations and the multinomial variance is not affected by this as the dispersion parameter is assumed to be one in both cases.

The approximated analytically derived value of the variance of each length proportion,  $\hat{p}_{l,y,pos}$ , was calculated using equation (1). These variances were compared with the simulated variances. Further, the estimated probabilities and the variances of these obtained by the traditional multinomial approach were calculated as well. These last two estimates do not take the variance between hauls (the overdispersion) into account.

### **Analysing age distributions for given length using continuation-ratio logits**

As the number-at-age in a sample also is an ordinal multinomial variable, the model may equally well be applied to age distributions. In this case, the observations in the conditional binomial distribution are the number of fish of age  $a$ , given the age is  $a$  or greater. The relationship between the length of the fish and the continuation-ratio logit of the probability of being of a given age in a length group has been shown analytically by Kvist (1999) to be simple in certain cases (results are summarised by the authors in the Appendix). Thus, if the length distribution of an age group follows a normal distribution, the relationship between length and the continuation-ratio logits can be approximated by a second-degree polynomial. A skewed length distribution such as a gamma distribution will lead to an approximately linear relationship between the continuation-ratio logits and length.

As the survey took place in early spring before the 0-group settled in the area, only the age groups 1, 2 and 3<sup>+</sup> were considered in the analyses. The primary aim of this model was to smooth the age distribution over lengths and not the age distribution at a given length. It was, therefore, considered to facilitate interpretation of the results to model age groups separately and thus not smooth over multinomial orders as described in the section on length distributions.

As the age determination of the two different laboratories may differ and slight changes in age structure may have taken place in the two weeks between the surveys, a laboratory

effect was included in the model. We assumed that this effect is independent of length, resulting in the full model

$$\log\left(\frac{\pi_{a|l,y,pos}}{1-\pi_{a|l,y,pos}}\right) = c_a + c_{a,y} + c_{a,pos} + c_{a,y,pos} + d_{a,lab} \\ + (g_a + g_{a,y} + g_{a,pos} + g_{a,y,pos}) * l \\ + (h_a + h_{a,y} + h_{a,pos} + h_{a,y,pos}) * l^2$$

where  $a$  denotes age,  $l$  denotes length,  $\pi_{a|l}$  denotes the conditional proportion of age  $a$  for given length,  $l$ , and  $lab$  denotes the laboratory at which the otoliths were read while  $c$ ,  $d$ ,  $g$  and  $h$  are parameters to be estimated in the model.

Estimation of parameters in the model is performed by minimising the deviance. However, when length groups in either end of the length spectre with an observed probability of 0 or 1 were included, convergence problems occurred due to the inability of the minimisation routine (SAS<sup>®</sup> GENMOD (SAS<sup>®</sup> Institute 1996)) to detect a global minimum in deviance (McCullagh and Nelder 1989). Therefore, these length groups were excluded from the data. This had virtually no effect on the estimated parameters and conditional probabilities while eliminating the problems in the minimisation procedure. This meant that fish smaller than 8 cm or greater than 16 cm were excluded from calculations in analysis of the probability of being age 1. Fish smaller than 11 cm were excluded in the analysis of the continuation-ratio logits of the 2-group.

### **Estimation of the length distribution by age**

The estimated length distribution for a given age group,  $\hat{p}_{l|a}$ , can be estimated by



$$(4) \quad \hat{p}_{l|a} = \frac{\hat{p}_l \hat{p}_{al}}{\sum_l \hat{p}_l \hat{p}_{al}}$$

where  $\hat{p}_{al}$  denotes the estimated probability of being  $a$  year old for a given length  $l$ , and  $\hat{p}_l$  denotes the estimated probability of being length  $l$ . As the length distribution may differ with year and position, a length distribution was estimated for each of the year/position combinations where more than 100 sandeel were measured. The average length distribution over positions of an age group in a given year was calculated as a weighted average, using the mean number caught at each position as weights.

## Results

### Length distributions

Observed length distributions, estimated multinomial length distribution and 95% confidence limits of this at two positions are displayed in Figure 2. The results of testing of the significance of the parameters in the full model of length distribution (equation (3)) show that the parameters  $b_n(y, pos)$  for  $n$  equal 6 and 7 were not significantly different from zero, which means that the continuation-ratio logit could be described by a fifth degree polynomial in length (Table 2). The fifth degree parameter,  $b_5$ , does not vary with year and position. The part of the polynomial which is of fourth degree or less varies with year, position or their crossed effect, indicating that the length distribution is not simply

changed towards larger or smaller fish at all positions from one year to the next. Note that the polynomial parameters estimated are not of particular interest but are merely tools used to model the fact that length distributions are usually smooth functions of length. The predicted length distributions for different positions are correlated and, therefore, there is a gain in precision of the predictions from incorporating this into the model as opposed to examining each position separately. The proportion of the total deviance explained by the model is high (93%). The rather high number of parameters estimated (91) should be compared with the multinomial alternative of estimating 271 probabilities (one for each length group at each year and position). The deviation between positions is greater than the deviation between years (15% as opposed to 8.8% of total deviance). It is interesting to note that the greater part of the variation between positions can be explained by a greater proportion of smaller or larger fish at some positions in both years, rather than a general trend towards larger fish in one of the years. The data were slightly overdispersed (dispersion parameter = 1.45).

The estimated conditional probabilities,  $\hat{\pi}_{l,y,pos}$ , the corresponding observed values and the 95% confidence limits of the estimates are plotted against length in Figure 3 for the two positions for which the multinomial length distribution is seen in Figure 2. Correspondingly, the unconditional length distributions predicted by the model for the same two positions are shown in Figure 4 along with the simulated confidence limits of this prediction. The confidence limits obtained by the suggested model are clearly smaller than the confidence limits of the multinomial probability usually employed (Fig. 2).

### **Comparison of simulated and approximated variance**

The length distributions estimated by applying the model to the simulated data did not differ significantly from the input length distributions from which the simulations were derived (the probability of difference in no case exceeded 0.50), indicating that the

method provides unbiased estimates of the length proportions. The relative difference between the standard deviation of the predicted length distribution approximated by equation (1) and the standard deviation calculated from the simulations is shown in Figure 5a. If the covariances assumed in equation (1) to be zero are in fact negative, then the approximated standard deviation will be an overestimate compared to the true value for all groups but the smallest. The relative difference between the variances is below 0.5 for the length groups 7 cm and 8 cm, but increases rapidly with length, indicating substantial negative correlation between the estimated conditional probabilities.

Comparing the simulated standard deviation of the model predictions with the standard deviation of the unsmoothed multinomial probabilities (Fig. 5b), the traditional method consistently provides estimates with higher standard deviation. The reduction in standard deviation ranges from a minimum of 29% on average for the smallest length group to a maximum of 118% on average for the 14 cm lengthgroup, indicating the advantage of utilising the knowledge of the smoothness of the catch as a function of length.

### **Age distribution for given length**

Leaving out insignificant effects, the model of the conditional probability of being age 1 became

$$\log\left(\frac{\pi_{1|l,y,pos}}{1 - \pi_{1|l,y,pos}}\right) = c_1 + c_{1,y} + c_{1,pos} + f_1 * l$$

There was thus no significant effect of length squared or laboratory. The reduced model describes the data well, explaining 91% of the total deviation (Table 3, Fig. 6). The

number of sandeel aged in each length group varies, with a mean of 16 for the lengths 8 to 12 and a mean of 3 for the lengths 13 to 16. The observed value of  $p = 1$  in Figure 6a at 14 cm is derived on the basis of 1 aged fish. Further, the logit of 0 and 1 are not defined, and these can therefore not be plotted in Figure 6b. However, these values are included in the model fitting process and the slope of the line therefore appears somewhat high.

By far the most important factor influencing the conditional probability is length (67%), demonstrating the importance of including this information in the model. The year effect is the second most important factor (20%). The same age length key should therefore not be applied in the two years. The lack of evidence of a second-degree effect of length indicates that the length distribution by age is either skewed (resembling a gamma distribution) or that the variances of the length distributions of age groups 1 and 2 are similar (Appendix). The percentage reduction in standard deviation obtained by using the model ranges from 53% (at 12 cm) to 247% (at 14 cm).

For age group 2, length once again explains more than half the variation in the probability ( $r^2 = 0.517$ ,  $P < 0.0001$ ). As for the 1-group, year effect is the second most important parameter with an  $r^2$  of 0.139. In contrast to the model of the conditional probability of being age 1, the reduced model showed significant lab effects ( $r^2 = 0.018$ ,  $P < 0.0002$ ), perhaps due to the increased difficulty in determining age of older fish. Including a significant crossed effect between year and length as well ( $r^2 = 0.010$ ,  $P < 0.0068$ ), the proportion of the total deviation explained is somewhat less than in the model of age 1.

### **Length distribution by age**

The length distributions estimated by equation (4) for 1- and 2-group sandeel in 1997 and 1998 are unimodal and smooth, and the distribution of the 2-year olds is slightly broader than the distribution of 1-group (Fig. 7). The distributions of the 1-group in both

years and the 2-group in 1997 are symmetrical and may be approximated by normal distributions. The distributions of 2-group in 1998 and the 3+-groups in both years are very skewed, but the distributions of the 2- and 3+-group in 1998 also show a tendency to bimodality. Whether this is caused by the actual length distribution or sampling problems such as problems with the accuracy of the age determinations this year is not possible to determine from this study.

## **Discussion**

The method suggested here has three main advantages compared to the standard use of the multinomial distribution. One is that entire multinomial distributions such as length distributions or age distributions by length may be modelled using generalised linear models in existing software. The theory of model selection, testing of hypotheses, estimation and prediction developed for these models may be used directly. Further, it is likely that the  $\chi^2$  distribution is a better approximation to the distribution of the deviance than the usual test in the multinomial distribution and that the power of the test of the new method is greater than the multinomial  $\chi^2$  test as the new method utilises the ordinal nature of the outcomes by simultaneously utilising data for all length groups contrary to the multinomial approach. However, the actual improvement in the example shown here has not been tested. Secondly, including the smoothness of the length distribution and the age-distribution as a function of length is straight-forward and improves the precision of the estimates. The suggested method therefore provides more accurate estimates of the length and age distribution than the multinomial approach. Estimates are furthermore unbiased. Thirdly, overdispersion is easily incorporated in the analyses allowing the data to vary more than expected in a traditional multinomial distribution. If the variance in the collected data is greater than expected in the multinomial case, this may alternatively be addressed by analysing compound multinomial distributions in which the probabilities of

the different outcomes are assumed to vary between samples (Smith and Maguire 1983) or by assuming the probabilities of the outcomes to be dirichlet distributed (Williams and Quinn 1998). However, neither of these methods incorporate the smoothness of the distributions, and they are both sensitive to the inadequacy of the  $\chi^2$  distribution in the small sample case.

We have not been able to find other examples in which generalised linear models and continuation-ratio logit analyses have been used simultaneously for all outcomes of a multinomial distribution thus enabling the relationship between the index of the group and the probability to be modelled as smooth. However, separate fitting for each index group as suggested by Agresti (1990) has been used in analyses of age distributions (Kvist et al. 2000).

The results of a model of the continuation-ratio logits are often more easily interpreted than a model of the corresponding multinomial distribution, as a positive factor in the continuation-ratio logit model may simply be interpreted as i. e. ‘increasing the probability of young fish’ in the case of analyses of age distributions. This was demonstrated in the reduced model of the conditional probability of being age 2, where one laboratory consistently aged a higher proportion of the fish as 2-group than the other laboratory. An additional advantage of the present approach compared to the traditional multinomial method is the ability to predict or interpolate length distributions or age compositions by length as well as the variances associated. This is particularly useful in cases where in a particular intermediate length group either no fish were caught, all fish were of a particular age or no fish were aged. Traditional methods are unable to estimate the proportion of a given age at this length. These problems may alternatively be approached by assuming the length distribution of the age groups to be normal (Labonté 1983; Gudmundsdóttir et al. 1988) or some other known distribution (Martin and Cook

1990). However, the method presented in this paper has the advantage of being able to describe unknown skewed or bimodal as well as normal length distributions by fitting the best possible polynomial in length to the logit, thereby letting the data dictate the age specific length distribution. When estimates from the models of length and age distribution are combined, they provide smooth length distributions of the age groups. The data are not smoothed to the extent where problems in the input data become invisible as is the case when length distributions at age are assumed to be normal. Further, length at age in the catch sample is sometimes known to deviate from normality, e.g. if catchability increases with length. The youngest age groups in a trawl survey catch will then have a skewed length distribution. Catchability problems, ageing problems, bimodal size at age distributions etc. are still detected when applying the method. In the simple case where length distributions of fish of age  $a$  and age  $a+2$  do not overlap, the relationship between the length distribution of the age group and the continuation-ratio logit model is simple as shown analytically by Kvist (1999) and summarised in the Appendix. In the data analysed here, the overlap between age group 1 and 3+ is indeed limited, and so the approximation is reasonable. However, slower growing species or species with large variations in growth rate may exhibit overlap between the length distributions of several age groups. The validity of the approximation of the continuation-ratio logit to a second degree polynomial in length in this case remains to be examined. Nevertheless, even if the approximation is not valid, a polynomial in length may still provide the best practical description of the continuation-ratio logit as a function of length.

When separate models of the continuation-ratio logit of the proportion at age by length are fitted, analytical calculation of the variance of the untransformed age proportions is simple. However, for length distributions in which the length dependent logits have been

analysed simultaneously, it is only possible to obtain poor approximations of the variance of the untransformed length proportions. The approximations may be appropriate for a small number of length groups, but Monte Carlo methods are needed to obtain reliable variance estimates for all length groups. Overdispersion should furthermore be taken into account when performing the simulations, if the dispersion parameter differs from 1. Thus, comparing mean lengths in one or more broader length groups by the  $\chi^2$ -test of homogeneity (Baird 1983; Zwanenburg and Smith 1983; Engås and Soldal 1992) may in many cases provide adequate estimates for which variances are more easily calculated.

The method presented here provides estimates of age composition in a sample with a higher precision than the traditional methods. The model further provides opportunities to test for the effect of factors such as time or area, which will enable the analyst to decide when to join length distributions or age-length keys and when to apply them separately to obtain the greatest accuracy in the results. In the past, extensive research has been designated to obtaining age-length keys with the highest precision at the lowest possible cost (Kimura 1977; Baird 1983; Lai 1993). Analysing data in the way suggested here may alter the optimal sampling strategy and may even allow the same precision to be obtained with fewer samples and thus lower cost. The implication of this for the optimal sampling strategy remains to be determined.

### **Acknowledgements**

The data on sandeel catches were collected and analysed as part of a project carried out with financial support from the Commission of the European Communities under study contract 95/078 and the UK Ministry of Agriculture, Fisheries and Food under contract CSA 4405. We thank Simon Greenstreet at the Institute of Fisheries Research Services, Marine Laboratory, UK, Scotland, for permission to use age-length observations collected



by the Scottish agency. Trine Kvist provided valuable help in developing the method and Morten Vinther and Poul Thyregod provided useful comments on an earlier draft.

## References

- Agresti, A. 1990. Categorical data analysis. Wiley series in probability and mathematical statistics pp. 306-346.
- Baird, J. W. 1983. A method to select optimum numbers for aging in a stratified random approach. *In* Sampling commercial catches of marine fish and invertebrates. *Edited by* W. G. Doubleday and D. Rivard. Can. Spec. Publ. Fish. Aquat. Sci. **66**: 161-164.
- Cotter, A. J. R. 1998. Method for estimating variability due to sampling of catches on a trawl survey. Can. J. Fish. Aquat. Sci. **55**(7): 1607-1617.
- Cramér, H. 1946. Mathematical methods of statistics. Princeton University Press.
- Davison, A. C., and Hinkley, D. V. 1997. Bootstrap methods and their application. Cambridge University Press.
- Engås, A., and Soldal, A. V. 1992. Diurnal variations in bottom trawl catch rates of cod and haddock and their influence on abundance indices. ICES J. Mar. Sci. **49**: 89-95.
- Fridriksson, A. 1934. On the calculation of agedistribution within a stock of cod by means of relatively few age determinations as a key to measurements on a large scale. Rapp. P. V. Reun., Cons. Int. Explor. Mer. **86**: 1-14.
- Gudmundsdóttir, Á., Steinarsson, B. Æ., and Stefánsson, G. 1988. A simulation procedure to evaluate the efficiency of some otolith and length sampling schemes. ICES C. M. 1988/D:14.
- Kimura, D. K. 1977. Statistical assessment of the age-length key. J. Fish. Res. Board Can. **34**: 317-324.
- Kvist, T. 1999. Statistical modelling of fish stocks. Ph. D. Thesis **64**, Institute of Mathematical Modelling, Danish Technical University, Lyngby, Denmark.

Kvist, T., Gislason, H., and Thyregod, P. 2000. Using continuation-ratio logits to analyze the variation of the age composition of fish catches. *J. App. Stat.* **27**(3): 303-320.

Labonté, S. S. M. 1983. Aging capelin: Enhancement of age-length keys and importance of such enhancement. *In Sampling commercial catches of marine fish and invertebrates. Edited by W. G. Doubleday and D. Rivard. Can. Spec. Publ. Fish. Aquat. Sci.* **66**: 171-177.

Lai, H. L. 1993. Optimal sampling design for using the age-length key to estimate age composition of a fish population. *Fish. Bull. U.S.* **91**: 382-388.

Martin, I., and Cook, R. M. 1990. Combined analysis of length and age-at-length data. *J. Cons. Int. Explor. Mer* **46**: 178-186.

Mathsoft Inc. 1997. S-Plus users guide. Data Analysis Products Division, Mathsoft Inc., Seattle, WA.

McCullough, P., and Nelder, J. A. 1989. Generalized linear models. Monographs on statistics and applied probability 37. Chapman and Hall, 2<sup>nd</sup> ed.

SAS Institute Inc. 1996. SAS/STAT<sup>®</sup> Software: Changes and enhancements through release 6.11, Cary, NC. SAS Institute Inc.: 231-317.

Schnute, J., and Fournier, D. 1980. A new approach to length-frequency analysis: Growth structure. *Can. J. Fish. Aquat. Sci.* **37**: 1337-1351.

Smith, S. J., and Maguire, J. J. 1983. Estimating the variance of length composition samples. *In Sampling commercial catches of marine fish and invertebrates. Edited by W. G. Doubleday and D. Rivard. Can. Spec. Publ. Fish. Aquat. Sci.* **66**: 165-170.

Williams, E. H., and Quinn, T. J. 1998. A parametric bootstrap of catch-age compositions using the dirichlet distribution. *In Fishery stock assessment models. Edited by Funk, F., Quinn II, T. J., Heifetz, J., Ianelli, J. N., Powers, J. E., Schweigert, J. F., Sullivan, P. J.*

and Zhang, C.-I. Alaska Sea Grant College Program Report No. AK-SG-98-01, University of Alaska Fairbanks: 371-384.

Winslade, P. 1974. Behavioural studies of the lesser sandeel *Ammodytes marinus* (Raitt) III. The effect of temperature on activity and the environmental control of the annual cycle of activity. *J. Fish. Biol.* **6**: 587-599.

Zwanenburg, K. C. T., and Smith, S. J. 1983. Comparison of finfish length-frequency distributions estimated from samples taken at sea and in port. *In* Sampling commercial catches of marine fish and invertebrates. *Edited by* W. G. Doubleday and D. Rivard. *Can. Spec. Publ. Fish. Aquat. Sci.* **66**: 189-193.

## Appendix

The relationship between the length and the continuation-ratio logit of the probability of being a particular age in the catch given the length can analytically be shown to be simple (Kvist 1999). A brief summary of the derivation of the relationship is given in the following.

The probability of a fish in the sample being age  $a$ , given the fish is of length  $l$ ,  $p_{a|l}$ , can be expressed by

$$p_{a|l} = \frac{p_{l|a} p_a}{\sum_a p_{l|a} p_a}$$

where  $p_{l|a}$  is the probability of a fish in the catch being length  $l$  given the age of the fish is  $a$  and  $p_a$  is the probability of a fish in the catch being age  $a$ . Note that this relationship is not affected by stratifying age-sampling by length, as long as age-samples are taken at random within each length group. The probability,  $p_{a|l}$ , is thus proportional to the length distribution at age  $a$  and the probability of being at age  $a$ .

The continuation-ratio logit for  $p_{a|l}$  is then

$$(A1) \quad \log \left( \frac{p_{a|l}}{\sum_{i \geq a+1} p_{i|l}} \right) = \log \left( \frac{p_{l|a} p_a}{\sum_{i \geq a+1} p_{l|i} p_i} \right)$$

If the length distribution at age,  $p_{l|a}$ , can be described by a normal distribution, equation (A1) can be written as

$$\log \left( \frac{p_{al}}{\sum_{i \geq a+1} p_{il}} \right) = \log p_a - \log \sigma_a - \frac{(l - \mu_a)^2}{2\sigma_a^2} - \log \left( \sum_{i \geq a+1} \frac{p_a}{\sigma_i} \exp \left( -\frac{(l - \mu_i)^2}{2\sigma_i^2} \right) \right)$$

where  $\mu_a$  and  $\sigma_a$  are mean and standard deviation of the length distribution of age group  $a$ .

This is unfortunately not a simple function of  $l$ . However, if age group  $a$  can be assumed to overlap only with age group  $a+1$  and not with age group  $a+2$ , the expression may be approximated by:

$$\log \left( \frac{p_{al}}{\sum_{i \geq a+1} p_{il}} \right) = \log p_a - \log \sigma_a - \frac{(l - \mu_a)^2}{2\sigma_a^2} - \log p_{a+1} + \log \sigma_{a+1} + \frac{(l - \mu_{a+1})^2}{2\sigma_{a+1}^2}$$

Thus, the continuation-ratio logit may be approximated by a second-degree polynomial in length, provided age group  $a$  does not overlap with age group  $a+2$ . Furthermore, if the standard deviations of age group  $a$  and age group  $a+1$  are equal, the continuation-ratio logit may be expressed by a first-degree polynomial in length.

The approximation is equally simple if the length distribution of the age groups is assumed to be gamma distributed with mean  $k_a\beta_a$  and variance  $k_a\beta_a^2$ , as the expression then becomes

$$\log\left(\frac{p_{a/l}}{\sum_{i \geq a+1} p_{il}}\right) = \log p_a - \log \Gamma(k_a) - k_a \log \beta_a + (k_a - 1) \log l - \frac{1}{\beta_a} \\ - \log p_{a+1} + \log \Gamma(k_{a+1}) + k_{a+1} \log \beta_{a+1} - (k_{a+1} - 1) \log l + \frac{1}{\beta_{a+1}}$$

where

$$\Gamma(k) = \int_0^{\infty} t^{k-1} \exp(-t) dt$$

and overlap between age group  $a$  and  $a+2$  is still assumed to zero. This expression is linear in  $\log(l)$  and is approximately linear in  $l$  in the range considered in the present paper.

Table 1. A multinomial data set with the ordinally scaled outcomes,  $i$ , and the corresponding data set for analysis in a generalised linear model of a binomially distributed variable with the explaining variable  $i$  and the observed conditional probabilities  $\pi_i$ .

Outcome ( $i$ )	Number of observations of $i$ ( $n_i$ )	Observed probability of $i$ ( $p_i$ )	Number of observations of index $i$ or greater ( $n_{i+}$ )	Conditional probability of $i$ ( $\pi_i$ )
1	10	0.19	54	0.19
2	25	0.46	44	0.57
3	14	0.26	19	0.74
4	3	0.06	5	0.60
5	2	0.04	(2) <sup>a</sup>	(1) <sup>a</sup>

<sup>a</sup>As the conditional probability of the highest outcome is always one, it is not included in the analyses.



Table 2. Generalised linear model of the continuation-ratio logit of the probability of

being of length  $l$ .  $f_1(l) = \sum_{n=0}^5 b_n l^n$ ,  $f_2(l, year) = \sum_{n=0}^3 b_{n,y} l^n$ ,  $f_3(l, position) = \sum_{n=0}^3 b_{n,pos} l^n$  and

$f_4(l, year, position) = \sum_{n=0}^3 b_{n,y,pos} l^n$ . See text for definition of  $b_n$ .

Source	Deviance	df	F (type 1)	Pr>F	$r^2$	Cum. $r^2$
Total	24,045	1036				
$f_1(l)$	13,901	5	1585	0.0001	0.578	0.578
$f_2(l,year)$	2120	4	292	0.0001	0.088	0.666
$f_3(l,position)$	3616	36	55	0.0001	0.150	0.817
$f_4(l,year,position)$	2654	36	41	0.0001	0.110	0.927
Position* $l^4$	74	9	4.7	0.0001	0.003	0.930
Year* $l^4$	1	1	0.8	0.3594	0.000	0.930
Year*position* $l^4$	13	9	0.8	0.629	0.001	0.931
Year* $l^5$	1	1	0.8	0.3608	0.000	0.931
Position* $l^5$	36	9	2.3	0.017 <sup>a</sup>	0.001	0.932
Year*position* $l^5$	9	9	0.6	0.831	0.000	0.933

<sup>a</sup> The parameters no longer have a significant effect when the model is reduced.

Table 3. Model of probability of being age 1.

Source	Deviance	df	F (type 1)	Pr>F	r <sup>2</sup>	Cum. r <sup>2</sup>
Total	2791	403	.	.	.	.
Length	1872	1	2225	0.0001	0.671	0.671
Year	549	1	653	0.0001	0.197	0.867
Position	108	9	14	0.0001	0.039	0.906
Lab	7	1	7.7	0.0059 <sup>a</sup>	0.003	0.909
Year*position	19	7	3.2	0.0027 <sup>a</sup>	0.007	0.915
Year*length	13	9	1.8	0.0660	0.005	0.920
Length <sup>2</sup>	1	1	1.0	0.3215	0.000	0.921
Year*length <sup>2</sup>	12	6	2.3	0.0356 <sup>a</sup>	0.004	0.925
Position*length	1	1	1.6	0.2133	0.000	0.925
Year*position*length	1	1	1.2	0.2666	0.000	0.926
Position*length <sup>2</sup>	4	9	0.6	0.8213	0.001	0.927
Year*position*length <sup>2</sup>	1	6	0.2	0.9878	0.000	0.927
Residual	203	335				

<sup>a</sup> The parameters no longer have a significant effect when the model is reduced.

## Figure texts

Fig. 1. Map of position of study area (a) and inset map of dredging positions 1-10 (b).

Fig. 2. Length distribution at position 1 (a) and 6 (b) in 1998. +: Observed probability. Solid line: Probability estimated by the multinomial distribution. Hatched lines: 95% confidence limits of the estimate (calculated by equation (2)).

Fig. 3. Probability of being a given length conditioned on being at least this length,  $\hat{\pi}_{l,y,pos}$ , at position 1 (a) and 6 (b) in 1998. +: Observed. Solid line: Predicted by model. Hatched lines: 95% confidence limits of the predicted mean.

Fig. 4. Length distribution at position 1 (a) and 6 (b) in 1998. +: Observed probability of being length 1. Solid line: Probability predicted by model. Hatched lines: 95% confidence limits of the mean (simulated).

Fig. 5. Relative difference in standard deviation (+) by length,

$\left( \frac{std_{estimated} - std_{simulated}}{std_{simulated}} \right)$ . The value of  $std_{estimated}$  calculated by a):

Assuming covariances to be zero (equation (1)) and b): The multinomial distribution (equation (2)). Line: Average relative difference.

Fig. 6. Probability of being age 1 (a) and the logit of this probability (b) as a function of length at position 30 in 1997. +: Observed values (calculated by haul). Solid line: Predicted value from model. Hatched line: 95% confidence limits of predicted value.

Fig. 7. Estimated length distribution of sandeel of age 1 (+), age 2 (O) and age 3 and older ( $\square$ ) in 1997 (a) and 1998 (b).

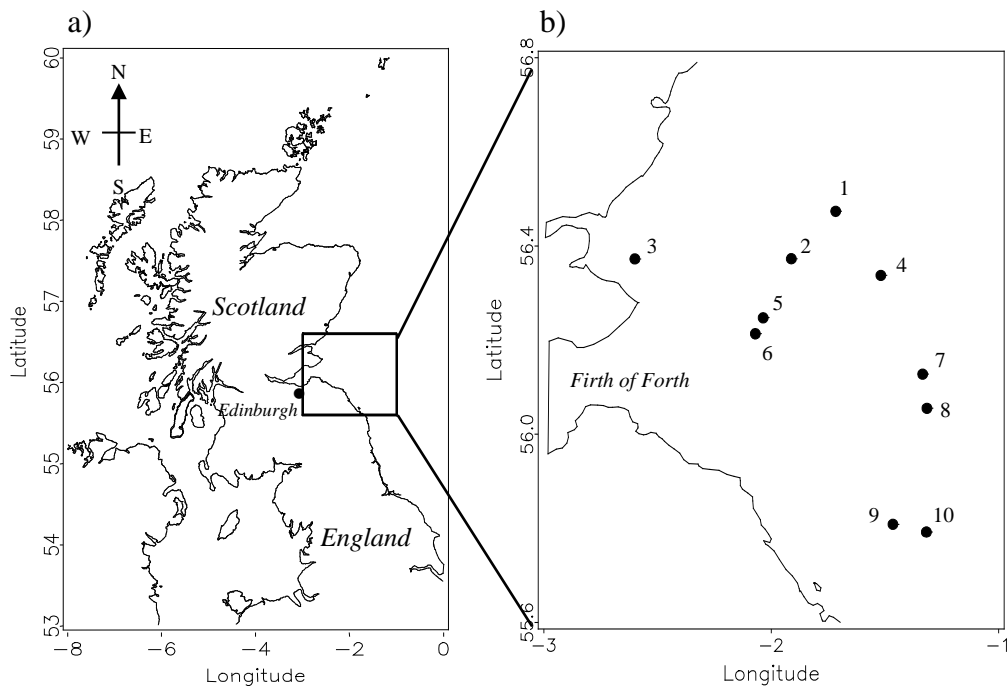


Fig. 1

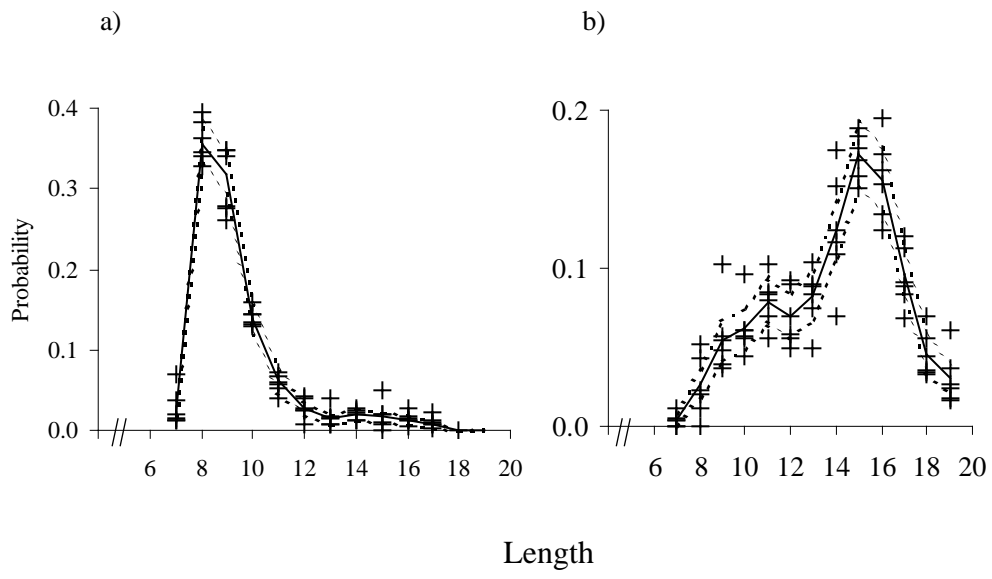


Figure 2.

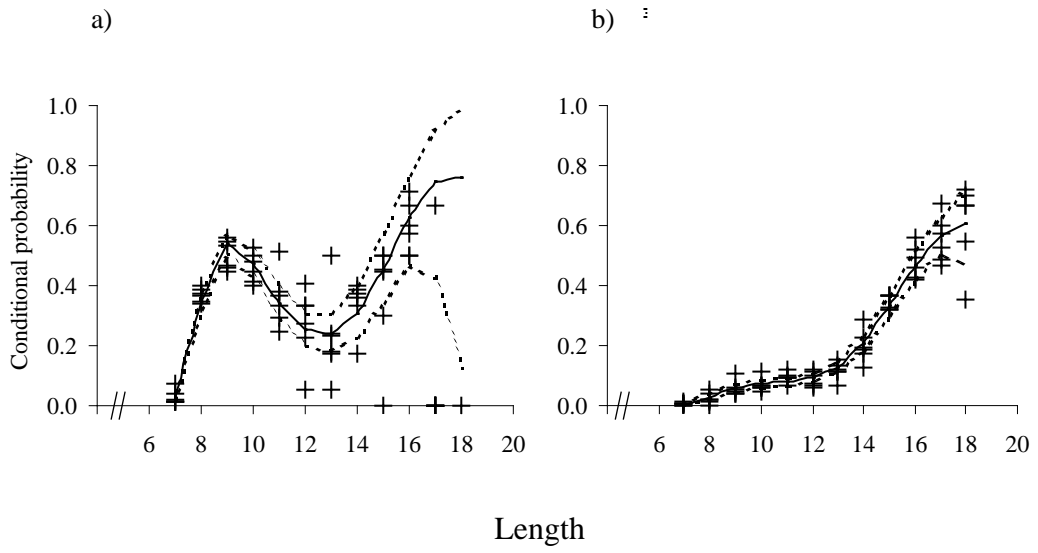


Figure 3

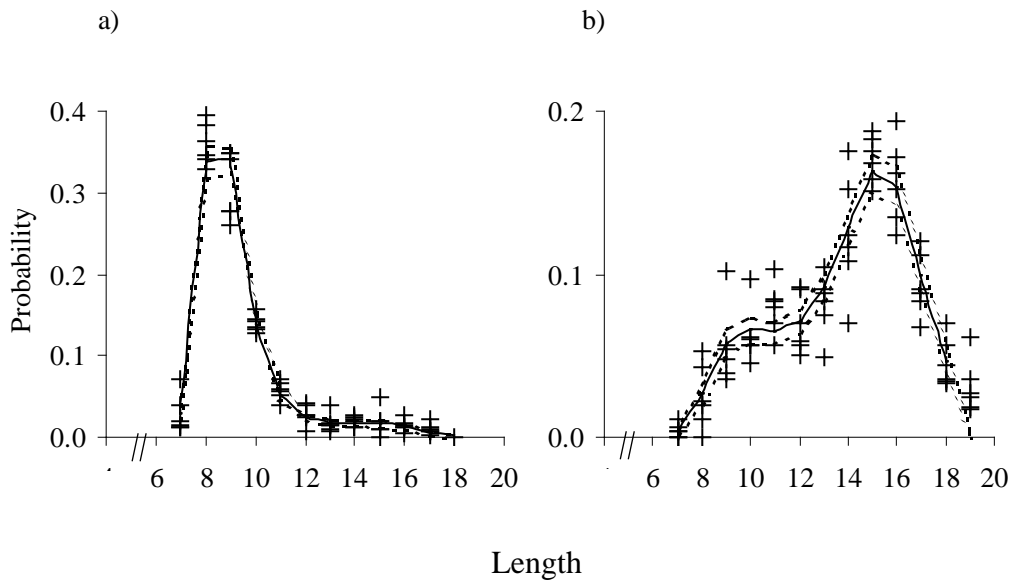


Figure 4



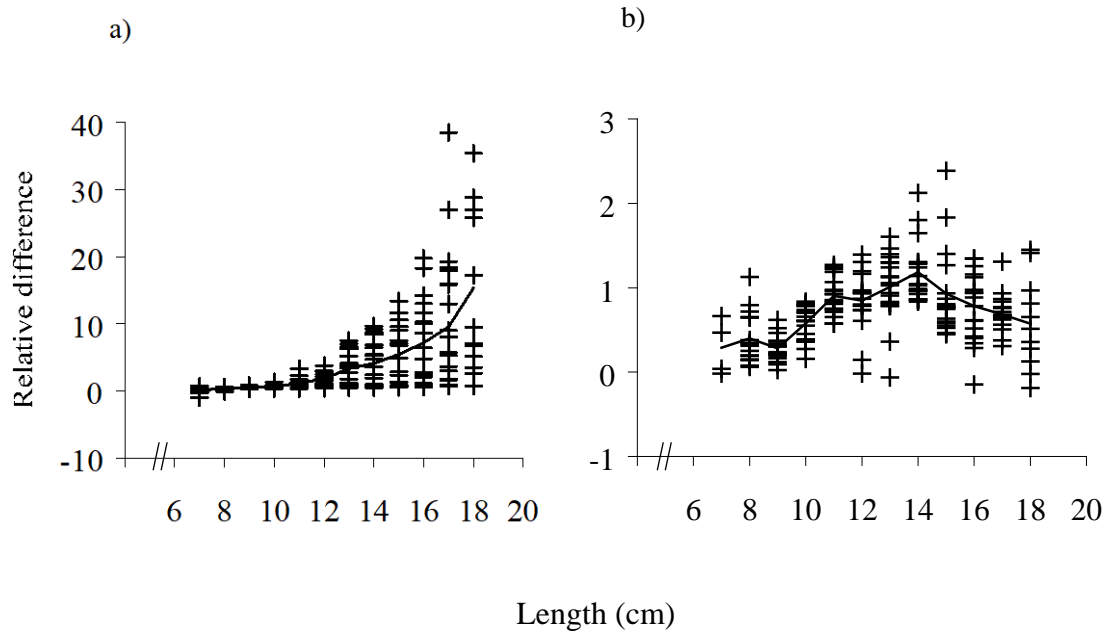


Figure 5

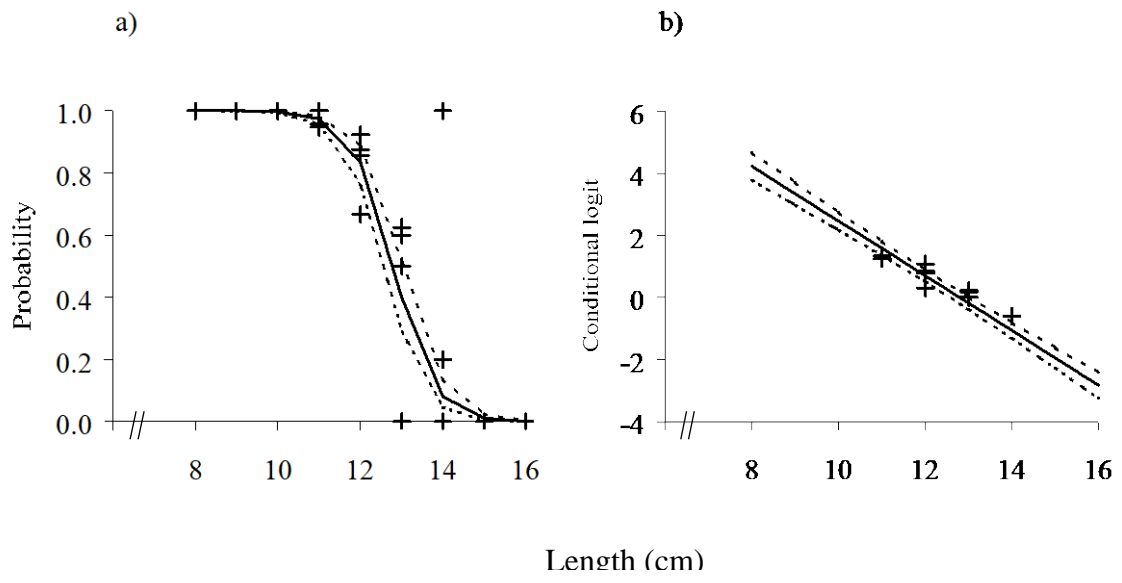


Figure 6