# Maximum likelihood estimates of mortality rates from catch at age data using survival analysis

**Marco Kienzle, Jason McGilvray, and You-Gan Wang**

**Abstract**: Survival analysis was applied to fisheries catch at age data to develop maximum likelihood estimators for stock assessment. This new method estimated natural mortality, fishing mortality and catchability from typical catch at age matrices. Monte Carlo simulations suggested estimates were unbiased and provided a better fit than the traditional multinomial likelihood. Application to a dataset from Queensland's sea mullet fishery (Australia) estimated natural mortality to be equal to $0.319 \pm 0.165$ year$^{-1}$.

*Key words:* survival analysis, maximum likelihood, age data, natural mortality, catchability, sea mullet.

## 1. Introduction

One purpose of stock assessment is to estimate mortality rates affecting fish stocks. This estimation problem is easier to solve for species that can be aged as opposed to those for which age can't be determined, as for example crustaceans. The reason is that mortality and longevity are inversely related, hence age is a measure of mortality. The central mortality model in fisheries research relating catch to the number of fish belonging to a cohort through time was proposed by Baranov (Quinn and Deriso, 1999). Given recruitment and mortality rates, the proportions of individuals at age in the catch can be calculated and used in a multinomial likelihood (Fournier and Archibald, 1982). This method has become by far the most common likelihood to integrate age data into modern stock assessment models (Francis, 2014; Maunder and Punt, 2013).

The deterministic exponential model in Baranov's catch equation has a statistical counterpart in the form of the exponential probability distribution function which first and second moments quantify the relationship between longevity and mortality rate (Cowan, 1998): the mean age of a cohort which abundance declines at a constant rate is the inverse of that rate. Adopting such a statistical view of the exponential decay of individual belonging to a cohort allowed to develop a set of maximum likelihood functions to estimate parameters of importance when assessing stocks. The field of survival analysis in statistics has created both a conceptual framework and refined methods to estimate mortality rates (Kleinbaum and Klein, 2005; Cox and Oakes, 1984) which are widely applied in the fields of medical research and engineering.

Despite the commonalities between survival analysis for medical and fisheries research, this theory has seldom been applied to animal ecology (Pollock et al., 1989): to our knowledge, there hasn't been any application to age data for the purpose of fisheries stock assessment. In this manuscript, we describe how to apply survival analysis to create likelihood functions of age data for the purpose of estimating natural and fishing mortality rates as well as gear selectivity. We started with a simplistic example using constant natural and fishing mortality rates to introduce fundamental concepts from survival analysis before moving to more sophisticated cases leading to its application to real data from the sea mullet fishery in Queensland (Australia). The proposed methods were tested with simulated data to characterize some of their properties and their capacity to estimate population dynamic parameters of interest. Finally, the application to the mullet fishery case study provided specific estimates of natural mortality, catchability and selectivity.

**Marco Kienzle.**[1] Queensland Dept of Agriculture, Fisheries and Forestry, Ecosciences Precinct, Joe Baker St, Dutton Park, Brisbane, QLD 4102, Australia;
University of Queensland, School of Agriculture and Food Sciences, St. Lucia, QLD 4072, Australia
**Jason McGilvray.** Queensland Dept of Agriculture, Fisheries and Forestry, Ecosciences Precinct, Joe Baker St, Dutton Park, Brisbane, QLD 4102, Australia
**You-Gan Wang.** University of Queensland, Centre for Applications in Natural Resource Mathematics, School of Mathematics and Physics, St. Lucia, QLD 4072, Australia

[1]Corresponding author (e-mail: Marco.Kienzle@daff.qld.gov.au).

## 2. Materials and methods

Fish can be assigned an age by examining its otolith, which is found just below its brain. Fish otoliths deposit calcium carbonate through time, thus increasing in size each year of their life. Microscopic observation of otolith sections often reveal alternate opaque and translucent zones, which can be used to assign individual fish to a particular age group.

Sampling programs in fisheries research centers around the world aim to collect a representative sample of fish each year to determine the distribution of age of any species of interest. In most cases, the data are binned into age-groups of width 1 year. For this reason, we split the lifespan of cohorts from their birth ($t \in [0; \infty]$) into $n$ yearly intervals from $a_1 = 0$ to the maximum age of $a_{n+1}$ years. While the theory presented in this document used that particular subdivision of time ($t$),

unequal ones also applies. In fact, an un-equal subdivision of time was used for the sea mullet case study.

### 2.1. The likelihood for constant natural and fishing mortality rates

The exponential decrease in abundance of individuals belonging to a single cohort due to constant natural ($M$) and fishing ($F$) mortality rates was described from a survival analysis point of view (Cox and Oakes, 1984) using a constant hazard function of time ($t$) and parameters $\theta$

$$h(t;\theta) = M + F \quad [1]$$

The probability density function (pdf) describing survival from natural and fishing mortality is

$$f(t;\theta) = (M+F)\, e^{-(M+F)t} \quad [2]$$

$$= \underbrace{M \times e^{-(M+F)t}}_{=f_1(t;\theta)} + \underbrace{F \times e^{-(M+F)t}}_{=f_2(t;\theta)} \quad [3]$$

Since age data belonging to individuals dying from natural causes are generally not available to fisheries scientists, we used only the component of the pdf that relates to fishing mortality ($f_2(t;\theta)$). This component of $f(t;\theta)$ integrates over the entire range of $t$ to

$$\int_{t=0}^{t=\infty} f_2(t;\theta)\, dt = \int_{t=0}^{t=\infty} F \times e^{-(M+F)t}\, dt \quad [4]$$

$$= \int_{t=0}^{t=\infty} f(t;\theta)\, dt - \int_{t=0}^{t=\infty} M \times e^{-(M+F)t}\, dt \quad [5]$$

$$= 1 - \int_{t=0}^{t=\infty} M \times e^{-(M+F)t}\, dt \quad [6]$$

$$= 1 - \frac{M}{M+F} \quad [7]$$

Hence, the pdf of catch at age data was obtained by normalizing $f_2(t;\theta)$

$$g(t;\theta) = \frac{1}{1 - \frac{M}{M+F}}\, f_2(t;\theta) \quad [8]$$

$$= \frac{M+F}{F}\, F \times e^{-(M+F)t} \quad [9]$$

$$= f(t;\theta) \quad [10]$$

The likelihood (Edwards, 1992) of a sample of fish caught in the fishery ($S_i$) was written as

$$\mathcal{L} = \prod_{i=1}^{n} \Big( \int_{t=a_i}^{t=a_{i+1}} f(t;\theta)\, dt \Big)^{S_i} \quad [11]$$

$$= \prod_{i=1}^{n} P_i^{S_i} \quad [12]$$

This is often referred to as the likelihood of a multinomial probability ($P_i$) where

$$P_i = \int_{t=a_i}^{t=a_{i+1}} f(t;\theta)\, dt \quad [13]$$

$$= e^{[-(M+F)a_i]} \big(1 - e^{[-(M+F)\times(a_{i+1}-a_i)]}\big) \quad [14]$$

is the probability of dying in the interval $[a_i; a_{i+1}]$.

The logarithm of the likelihood was

$$\log(\mathcal{L}) = \sum_{i=1}^{n} S_i \log\Big( \int_{t=a_i}^{t=a_{i+1}} f(t;\theta)\, dt \Big) \quad [15]$$

$$= \sum_{i=1}^{n} S_i \log\Big( \int_{t=a_i}^{t=a_{i+1}} (M+F)\, e^{-(M+F)t}\, dt \Big) \quad [16]$$

$$= \sum_{i=1}^{n} S_i \log\big( e^{-(M+F)\times a_i} - e^{-(M+F)\times a_{i+1}} \big) \quad [17]$$

[18]

In cases where sampling does not cover the entire range of age ($i \notin in[1;n], i \in [a;b]$ with $a > 1$ and $b < n$) for example because younger fish live outside the sampling area, the distribution is truncated ($\sum_i P_i < 1$) and using Eq. 18 would lead to erroneous estimations. In such cases, the relative proportions $p_i = P_i / \sum_{i=a}^{b} P_i$ are to be used in the likelihood function. Conditional on the total sample $S$ with age between age $a_i$ and $a_{i+1}$, with $i \in [a;b]$, the age frequency of the total sample $S$ follows the following multinominal distribution (up to a constant $S!/S_a!\ ...\ S_b!$) (see Wang (1999))

$$\mathcal{L} = \prod_{i=a}^{b} p_i^{S_i}. \quad [19]$$

This development illustrated an application of survival analysis to estimate mortality rates affecting a cohort of fish by maximum-likelihood using a sample of catch at age. This method was implemented in R (R Core Team, 2013) in the package Survival Analysis for Fisheries Research (SAFR) provided as supplement material. Numerical application were made available using the following commands: **library(SAFR); example(llfunc1)**.

Natural and fishing mortality cannot be disentangled with catch data only but the next section will show that the provision of effort data allowed to estimate both catchability ($q$) and natural mortality.

### 2.2. Estimating catchability and natural mortality

In this section, we assumed that a time series of effort ($E_i$) associated with a sample of catch at age ($S_i$) was available to the researcher. And the assumption that fishing mortality varied according to fishing effort through constant catchability ($q$) held: $F(t) = q\, E(t)$. In this situation, the hazard function was written as

$$h(t,\theta) = M + q\, E(t) \quad [20]$$

And the pdf

$$f(t,\theta) = (M + q\,E(t))\; e^{-Mt - q\int_0^t E(t)\;dt}$$

$$= \underbrace{M \times e^{-Mt - q\int_0^t E(t)\;dt}}_{=f_1(t;\theta)} + \underbrace{q\,E(t) \times e^{-Mt - q\int_0^t E(t)\;dt}}_{=f_2(t;\theta)}$$

As in the previous section, we had

$$\int_{t=0}^{t=\infty} f_2(t;\theta)\;dt =$$

$$1 - \int_{t=0}^{t=\infty} M \times e^{-Mt - q\int_0^t E(t)\;dt}\;dt$$

But we did not know an analytic solution to the integral since the function $E(t)$ was not specified. Nevertheless, as we knew the value of effort in any given interval ($\int_{t=a_i}^{t=a_{i+1}} E(t)\;dt = E_i = \int_{t=0}^{t=a_{i+1}} E(t)\;dt - \int_{t=0}^{t=a_i} E(t)\;dt, \forall i \in [1;n]$), we could calculate the value of $\int_{t=0}^{t=\infty} f_2(t;\theta)\;dt$ assuming $E(t)$ was constant over each interval $i$

$$\int_{t=0}^{t=\infty} f_2(t;\theta) =$$

$$1 - \sum_{i=1}^{n}\Big[-\frac{M}{M+q\,E_i}e^{-Mt-q\int_0^t E(t)\;dt}\Big]_{t=a_i}^{t=a_{i+1}} =$$

$$1 - \sum_{i=1}^{n}\frac{M}{M+q\,E_i}\big(e^{-M\,a_i - q\int_0^{a_i} E(t)\;dt} - e^{-M\,a_{i+1} - q\int_0^{a_{i+1}} E(t)\;dt}\big) =$$

$$\sum_{i=1}^{n}\frac{q\,E_i}{M+q\,E_i}\big(e^{-M\,a_i - q\int_0^{a_i} E(t)\;dt} - e^{-M\,a_{i+1} - q\int_0^{a_{i+1}} E(t)\;dt}\big)$$

In practice, $0 \leq \int_{t=0}^{t=\infty} f_2(t;\theta) \leq 1$ and took a specific value depending on the values of $M$, $q$ and $E_i$. Naming this constant value $K$, we could write the pdf of catch at age given that effort data are available as

$$g(t;\theta) = \frac{1}{K}\,f_2(t;\theta)$$

And the log-likelihood:

$$\log(\mathcal{L}) = \sum_{i=1}^{n} S_i \log\Big(\int_{t=a_i}^{t=a_{i+1}} g(t;\theta)\;dt\Big)$$

Numerical application of this method were made available using the following commands: **library(SAFR); example(llfunc2)**

Accounting for age-specific gear selectivity ($s(t)$) effects on fishing mortality ($F(t) = q\;s(t)\;E(t)$) was included in a similar way into the likelihood using constant value for selectivity at age. In practice, it is difficult to estimate $n$ additional selectivity parameters using only the data from a single cohort but processing several cohorts at the same time and assuming separability of fishing mortality rendered estimation of catchability, natural mortality and selectivity possible.

## 2.3. Estimates from catch at age matrix using fishing mortality separability

This section describes an application of survival analysis to matrices of catch at age, developed for the purpose of estimating catchability ($q$), selectivity at age ($s(t)$) and constant natural mortality ($M$). The matrix ($S_{i,j}$) containing a sample of fishes aged to belong to a particular age-group $j$ in year $i$ contains $n + p - 1$ cohorts. These cohorts were indexed by convention using $k$ ($k \in [1, n+p-1]$) and an increasing number $r_k$ ($1 \leq r_k \leq \min(n,p)$) identifying incrementally each age-group (see appendix p. **??** for more information). Each matrix $S_{i,j}$ has two cohorts with only 1 age-group representing them.

The derivation for a single cohort were the same as those presented in the previous section, here reproduced with indexations relative to a single cohort and accounting for selectivity

$$g_k(t;\theta) = \frac{q\;s(t)\;E(t) \times e^{-Mt - q\int_0^t s(t)}}{\sum_{l=1}^{r_k}\frac{q\;s_{k,l}\;E_{k,l}}{M+q\;s_{k,l}\;E_{k,l}}\big(e^{-M\;a_{k,l} - q\int_0^{a_{k,l}} s(t)\;E(t)\;dt} -}$$

The likelihood function of a catch at age matrix was build using each pdf specific to each cohort ($g_k(t;\theta)$):

$$\mathcal{L} = \prod_{k=1}^{n+p-1}\prod_{l=1}^{r_k}\Big(\int_{t=a_{k,l}}^{t=a_{k,l+1}} g_k(t;\theta)\;dt\Big)^{S_{k,l}}$$

The expression above is equivalent to

$$\mathcal{L} = \prod_{i,j} P_{i,j}^{S_{i,j}}$$

where the $P_{i,j}$ are the probabilities of observing a fish of a given age $j$ in year $i$ given by the hazard model. In this likelihood, the $P_{i,j}$ sum to 1 along the cohort instead of summing to 1 for each year as described for the multinomial likelihood in Fournier and Archibald (1982).

This method was implemented in R (R Core Team, 2013) in the package Survival Analysis for Fisheries Research (SAFR). Numerical application of this method are available using the following commands: **library(SAFR); example(llfunc3); example(llfunc4); example(llfunc5);**.

## 2.4. Monte Carlo simulations

Methods to estimate mortality and selectivity from a matrix containing a sample of number at age were tested with simulated datasets to characterize their performance. Variable number of cohorts ($n + p - 1 = 25$, 35 or 45); maximum age ($p = 8$, 12 or 16 years) and sample size of age measurement in each year varying from 125 to 2000 increasing successively by a factor 2. The simulated datasets were created by generating an age-structure population using random recruitment for each cohort, random constant natural mortality, random catchability and random fishing effort in each year (Tab. **??**). A catch at

age matrix was calculated using a logistic gear selectivity with 2 parameters:

$$[36] \qquad s_{a_i} = \frac{1}{1 + \exp(\alpha - \beta \times a_i)}$$

Several sampling strategies were implemented to assess how it affected mortality estimates. To test estimators derived from survival analysis, one would like to draw randomly from the probability distribution. This is obviously impossible in the real world because field biologists never have in front of them a entire cohort to chose from. Nevertheless, we implemented a sampling strategy (sampling strategy 1) that randomly selected from the entire simulated catch at age dataset as a benchmark. In the real world, samples can be drawn by accessing only a single year-class of every cohort every year, so the second strategy implemented was to simulate a random selection of a fixed number of sample ($N$) each year (sampling strategy 2). Finally, the third strategy investigated was to apply a weighting by the estimated total catch at age ($\hat{C}_{i,j}$) to the sample of number at age in the sample ($S_{i,j}$) – sampling strategy with weighting :

$$[37] \qquad \hat{C}_{i,j} = p_{i,j} \odot C_i \otimes v(j)$$

where $p_{i,j}$ is the proportion at age (see appendix p. **??**), $C_i$ is a column vector containing the total number of fish caught in each year $i$ and $v(j)$ is a row vector of 1's. And a weighted sample ($S^*_{i,j}$) was obtained using the fraction of total catch sampled

$$[38] \qquad S^*_{i,j} = \hat{C}_{i,j} \times \frac{\sum_{i,j} S_{i,j}}{\sum_i C_i}$$

Note that $\sum_{i,j} S_{i,j} = \sum_{i,j} S^*_{i,j}$.

Comparison with the multinomial likelihood proposed by Fournier and Archibald (1982) were made using differences in negative log-likelihood between that method and the survival analysis approach described in the present article. Simulated catch were used to calculate the proportion of individual at age, constraining them to sum to 1 in each year. This method to calculate proportions for the multinomial likelihood was regarded as the best case scenario because we expect any estimation algorithm based on the multinomial likelihood to, at best, match exactly the simulated catch at age. The logarithm of these proportions were then multiplied by the simulated age sample (weighted or not depending on the case) to calculate the log-likelihood as described in Fournier and Archibald (1982). This quantity was compared to that calculated using the survival analysis approach to determine which model best fitted the simulated data. This comparison ignored the number of parameter used in each model as the Akaike criteria would. The multinomial likelihood requires $n + p - 1$ more parameters to be estimated than the survival analysis because the former requires an estimate of recruitment for each cohort in order to calculate the proportion at age in the catch.

### 2.5. A case study: Queensland's sea mullet fishery

The straddling Sea Mullet (*Mugil cephalus*) population stretches along the east coast of Australia, with most landings occurring between 19$^o$S (approx. Townsville) and 37$^o$S (roughly the border between New South Wales and Victoria). Following recommendations from Bell et al. (2005), an existing (1999–2004) scientific survey design was modified from 2007 onward to include both estuaries and ocean habitats in order to provide representative demographic statistics for Queensland component of this fishery. The number of fish at age obtained by otolithometry (Tab. **??**) were analyzed to estimate natural mortality, catchability and gear selectivity.

Sea Mullet are thought to spawn in oceanic waters adjacent to ocean beaches from May to August each year. By convention, the birth date was assumed to be on July $1^{st}$ each year. Opaque zones are thought to be deposited on the otolith margin during spring through early summer (September to December). Biologists have come to the conclusion that the first identifiable opaque zone is formed 14 to 18 months after birth, and all subsequent opaque zones are then formed at a yearly schedule (Smith and Deguara, 2003). Each fish in the sample was assigned an age-group based on opaque zone counts and the amount of translucent material at the margin of otolith. Age-group 0–1 comprised fish up to 18 months old ($a_1 = 18$ months) while all subsequent age-groups spanned 12 months ($a_2 = 30$ months, $a_3 = 42$ months, etc ...).

Sensitivity of survival analysis estimates to these data, a matrix containing 7 years and 16 age-groups, were performed by truncating the dataset in 2 ways to assess the robustness of the method to varying number of years and age-groups. The first truncation removed the last and last-two years of data to evaluate the sensitivity of parameters estimates to addition/omission of data in order to anticipate possible effects of future addition of newly available data. The second truncation removed older age-groups from 10–11 to 15–16 to evaluate the importance of few old fish on natural mortality estimates as one could think *a priori* that these longer-lived individuals provided a lot of information on mortality.

## References

P.A. Bell, M.F. O'Neill, G.M. Leigh, A.J. Courtney, and S.L. Peel. Stock assessment of the Queensland-New South Wales sea mullet fishery (*Mugil cephalus*). Technical Report QI05033, Queensland Government, 2005.

G. Cowan. *Statistical Data Analysis*. Oxford Science Publications, 1998.

D.R. Cox and D. Oakes. *Analysis of survival data*. Chapman and Hall, 1984.

A.W.F. Edwards. *Likelihood*. Johns Hopkins University Press, 1992.

D. Fournier and C.P. Archibald. A General Theory for Analyzing Catch at Age Data. *Canadian Journal of Fisheries and Aquatic Sciences*, 39(8):1195–1207, 1982.

R.I.C.C. Francis. Replacing the multinomial in stock assessment models: A first step. *Fisheries Research*, 151(0):70–84, 2014.

D.G. Kleinbaum and M. Klein. *Survival Analysis: A Self-Learning Text*. Springer, 2005.

N.C. Krück, D.I. Innes, and J.R. Ovenden. New SNPs for population genetic analysis reveal possible cryptic speciation of eastern australian sea mullet (*Mugil cephalus*). *Molecular Ecology Resources*, 13(4):715–725, 2013.

R.J.G. Lester, S.E. Rawlinson, and L.C. Weaver. Movement of sea mullet mugil cephalus as indicated by a parasite. *Fisheries Research*, 96(23):129 – 132, 2009.

M.N. Maunder and A.E. Punt. A review of integrated analysis in fisheries stock assessment. *Fisheries Research*, 142:61–74, 2013.

K.H. Pollock, S.R. Winterstein, and M.J. Conroy. Estimation and analysis of survival distributions for radio-tagged animals. *Biometrics*, 45(1):pp. 99–109, 1989.

T. J. Quinn and R. B. Deriso. *Quantitative Fish Dynamics*. Oxford University Press, 1999.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`.

K. Smith and K. Deguara. Formation and annual periodicity of opaque zones in sagittal otoliths of *Mugil cephalus* (Pisces: Mugilidae). *Marine & Freshwater Research*, 54: 57–67, 2003.

Y.G. Wang. A maximum-likelihood method for estimating natural mortality and catchability coefficient from catch and effort data. *Marine & Freshwater Research*, 50:307–11, 1999.