CCAMLR Science, Vol. 1: 203-207

ESTIMATING CONFIDENCE INTERVALS FOR FISH STOCK ABUNDANCE ESTIMATES FROM TRAWL SURVEYS

W.K. de la Mare
Australian Antarctic Division
Channel Highway, Kingston 7050
Tasmania, Australia

Abstract

A method is developed for calculating asymptotic confidence intervals for estimates of abundance obtained from trawl surveys conducted by means of the swept area method, using likelihood ratios from Aitchison's delta distribution. Simulation tests of the method show that unbiased estimates of density and biomass can be obtained and that the estimated confidence intervals have close to the nominal coverage probability. Performance deteriorates in cases where few of the hauls contain fish and the coefficient of variation (CV) is high. The upper confidence bound appears to be slightly less reliable than the lower.

Résumé

Développement d'une méthode de calcul des intervalles de confiance asymptotiques des estimations d'abondance provenant des campagnes d'évaluation par chalutages menées au moyen de la méthode de l'aire balayée et en utilisant les rapports de probabilité de la distribution delta de Aitchison. Les tests par simulation de la méthode indiquent qu'il est possible d'obtenir des estimations non biaisées de densité et de biomasse et que les intervalles de confiance estimés ont une probabilité proche de la couverture nominale. La performance est moins bonne lorsque le nombre de traits contenant des poissons est faible alors que le coefficient de variation (CV) est élevé. Il semblerait que la valeur supérieure de l'intervalle ne soit pas aussi fiable que la limite inférieure.

Резюме

Разработан метод вычисления асимптотических доверительных интервалов по оценкам численности, полученным в результате траловых съемок, проведенных в соответствии с методом протраленных площадей, использующий соотношения вероятности по дельта-распределению Эйтчисона. Результаты имитационного испытания метода показывают, что можно получить несмещенные оценки плотности и биомассы, и что рассчитанные доверительные интервалы близки к вероятности номинального охвата. Данный метод становится менее эффективным тогда, когда рыба попадается лишь в небольшое количество тралений и коэффициент вариации (CV) высок. Представляется, что верхний доверительный предел менее надежен, чем нижний.

Resumen

Se elabora un método para calcular los intervalos asintóticos de confianza para las estimaciones de abundancia obtenidas de las prospecciones de arrastre mediante el método de área barrida, empleando razones de probabilidad de la distribución delta de Aitchison. Las pasadas de simulación del método indicaron que se pueden obtener cálculos correctos de densidad y de biomasa y que los intervalos de confianza calculados se acercan a la probabilidad de cobertura nominal. El rendimiento deteriora en casos en donde sólo algunos lances contienen peces y el coeficiente de variación (CV) es alto. El límite de confianza superior podría ser ligeramente menos fidedigno que el valor inferior.

Keywords: fish, trawl, survey, population, modelling, density, biomass, confidence intervals, CCAMLR

INTRODUCTION

In 1991 and 1992, the Working Group on Fish Stock Assessment (WG-FSA) drew attention, as a matter of priority, to the difficulties which had been experienced in the application of the swept area method for estimating fish stock abundance (Saville, 1977) and associated *t*-statistics, as a basis for confidence interval estimation, to species with patchy distributions, such as *Champsocephalus gunnari*. Although a workshop was held on this matter in Hamburg in 1992 (SC-CAMLR, 1992) progress on the statistical considerations of the swept area method was hampered by the absence of statisticians. This paper addresses some of the statistical issues which are of concern in the analysis of trawl surveys, and in particular, develops an improved method for the estimation of confidence intervals. This is a brief overview of some of the results to be presented in a more substantial paper on this subject which is currently in preparation.

BASIC METHODOLOGY

The statistical distribution of net haul densities has to allow for an often substantial probability that a given haul will produce a zero density estimate (i.e., the net is empty). The statistics of such distributions have been examined by Aitchison (1955), and Pennington (1983) has recommended using Aitchison's delta distribution as the underlying statistical model when analysing net haul survey data. This recommendation is followed in the method developed here. The delta distribution consists of a discrete probability at the origin and a lognormal distribution for the non-zero observations. The delta distribution has the following probability function:

$$\mathrm{f}\left(x;p,\lambda,\kappa^2\right)= (1-p)\mathrm{I}_0[x]+p\frac{1}{x\sqrt{2\pi}\kappa}\mathrm{e}^{-\frac{1}{2}\left(\frac{\ln x-\lambda}{\kappa}\right)^2}\mathrm{I}_{(0,\infty)}[x] \qquad (1)$$

where p is the proportion of observations for which $x > 0$, λ and κ^2 are the parameters of the lognormal distribution of the non-zero observations, $\mathrm{I}_0[x]$ is an indicator function which takes the value 1 when $x = 0$ and 0 otherwise, and $\mathrm{I}_{(0,\infty)}[x]$ takes the value 0 when $x = 0$ and 1 when $x > 0$. The first term represents a discrete probability mass at the origin and the second term, a probability density. If $p = 1$, (1) is the probability density function for a lognormal distribution. The statistical methodology which follows is also valid for the lognormal case. The log-likelihood of a vector of observations $x_1 \ldots x_N$ from a delta distribution is given by:

$$\ln\left[\wp\left(x_1\ldots x_N;p,\lambda,\kappa^2\right)\right]= (N-m)\ln(1-p)+m\ln p-\frac{m}{2}\ln\kappa^2 -\frac{1}{2\kappa^2}\sum_{x>0}\left(\ln x_i-\lambda\right)^2-\sum_{x>0}\ln x_i-\frac{m}{2}\ln 2\pi \qquad (2)$$

where N is the total number of observations and m is the number of non-zero observations. The last two terms are additive constants which can be ignored when maximising the likelihood function to calculate estimates. In the method described here, it is the densities in each haul which constitute the x_i. Using Aitchison's (1955) formulae, the minimum variance unbiased estimate of the mean density is:

$$\begin{aligned} \bar{d} &= \frac{m}{N}e^{\bar{y}}\mathrm{G}_m\left(\frac{1}{2}s^2\right), && m>1 \\ \bar{d} &= \frac{x_1}{N}, && m=1,\ x_1>0 \\ \bar{d} &= 0, && m=0 \end{aligned} \qquad (3)$$

where $\bar{y}$ and s^2 are the sample mean and sample variance (the unbiased estimate of the population variance) of the log of the non-zero observations and:

$$\mathrm{G}_m(t)=1+\frac{m-1}{m}t +\sum_{r=2}^{\infty}\frac{(m-1)^{2r-1}}{m^r(m+1)(m+3)\ldots(m+2r-3)}\cdot\frac{t^1}{r!} \qquad (4)$$

Pennington (1983) gives the following unbiased estimator for the variance of the mean density:

$$\begin{aligned} \hat{\sigma}^2 &= \frac{m}{N}\exp(2\bar{y}) \left\{\frac{m}{N}\mathrm{G}_m\left(\frac{1}{2}s^2\right)-\left(\frac{m-1}{N-1}\right)\mathrm{G}_m\left(\frac{m-2}{m-1}s^2\right)\right\}, && m>1 \\ \hat{\sigma}^2 &= \left(\frac{x_1}{N}\right)^2, && m=1 \\ \hat{\sigma}^2 &= 0, && m=0 \end{aligned} \qquad (5)$$

The usual sample statistics for the mean and variance are not minimum variance estimators for this distribution; the statistics of equations (3) and (5) are more efficient estimators (have lower expected standard errors).

Using a likelihood ratio approach (Cox and Hinkley, 1974), asymptotic confidence intervals on the mean density can be found as the roots of the following function:

$$q(d) = \left[\ln \mathcal{L}\left(\mathbf{x}; p, \lambda, \kappa^2\right) \middle| \begin{array}{l} p = \dfrac{m}{N}, \\ \lambda = \dfrac{1}{m}\displaystyle\sum_{x_i>0} \ln x_i, \\ \kappa^2 = \dfrac{1}{m}\displaystyle\sum_{x_i>0} \left(\ln x_i - \lambda\right)^2 \end{array} \right]$$

$$-\mathrm{Sup}\left[\ln \mathcal{L}\left(\mathbf{x}; p, \lambda, \kappa^2\right) \middle| \begin{array}{l} 0 < p \le 1, \\ \lambda = \ln\left(d / p G_m\left(\kappa^2/2\right)\right), \\ 0 < \kappa^2 < \infty \end{array} \right] - \frac{1}{2}\chi_{1,\alpha}{}^2 \qquad (6)$$

where $\mathbf{x}$ is the vector of observations $x_1 \ldots x_N$, $\chi^2_{1,\alpha}$ is the critical value of the χ^2 distribution with one degree of freedom, at the α probability level.

Abundance estimates (b) and their confidence intervals are calculated in the usual way as the product of the corresponding densities and the area covered by the survey (A) for example:

$$b = \bar{d}A \qquad (7)$$

The variance of the abundance estimate is the product variance of the density estimate and the square of the area of the survey:

$$\hat{\sigma}_b{}^2 = \hat{\sigma}^2 A^2 \qquad (8)$$

STRATIFIED SURVEYS

In the case of stratified surveys, it is assumed that the hauls from each stratum are drawn randomly from different, independent underlying delta distributions. In this case, the mean density for each stratum is calculated by applying the delta method estimator, equation (3). The total abundance estimate B is given by:

$$B = \sum_{j=1}^{k} \bar{d}_j A_j \qquad (9)$$

where $\bar{d}_j$ is the mean density in stratum j, and A_j is its area. The variance of B is given by:

$$\hat{\sigma}_B{}^2 = \sum_{j=1}^{k} \hat{\sigma}_j{}^2 A_j{}^2 \qquad (10)$$

An asymptotic confidence interval for the total abundance can be found from the log-likelihood expressed as a function of the total biomass, given by:

$$B = \sum_{j=1}^{k} b_j \qquad (11)$$

where b_j is the biomass estimate for stratum j. In calculating an asymptotic confidence interval we need to find the value of B which gives a specified critical value for the log likelihood. However, the likelihood is a function of the vector $\mathbf{b}$, so that the log-likelihood has to be maximised over the vector $\mathbf{b}$, subject to the constraint given in equation (11). This is achieved by maximising the log-likelihood over k - 1 of the stratum biomasses, with the remaining value fixed as:

$$b_1 = B - \sum_{j=2}^{k} b_j \qquad (12)$$

Thus, the confidence interval is given by the roots of the following function of B:

$$R\left(B; b_2 \ldots b_k\right) = \sum_{j=1}^{k} \left[\ln \mathcal{L}\left(\mathbf{x}_j; p_j, \lambda_j, \kappa_j{}^2\right) \middle| \begin{array}{l} p_j = \dfrac{m_j}{N_j}, \\ \lambda_j = \dfrac{1}{m_j}\displaystyle\sum_{x_{ij}>0} \ln x_{ij}, \\ \kappa_j{}^2 = \dfrac{1}{m_j}\displaystyle\sum_{x_{ij}>0} \left(\ln x_{ij} - \lambda_j\right)^2 \end{array} \right]$$

$$-\mathrm{Sup}\left[\sum_{j=1}^{k} \left(\ln \mathcal{L}\left(\mathbf{x}_j; p_j, \lambda_j, \kappa_j{}^2\right) \middle| \begin{array}{l} 0 < p_j < 1, \\ \lambda_j = \ln\left(\dfrac{b_j}{A_j p_j G_{mj}\left(\frac{1}{2}\kappa_j{}^2\right)} \right), \\ 0 < \kappa_j{}^2 < \infty \end{array} \right) \right] - \frac{1}{2}\chi_{1,\alpha}{}^2 \qquad (13)$$

where $\mathbf{x}_j$ is the vector of observations for stratum j, i.e. $x_{1,j} \ldots x_{Nj,j}$. The vectors of biomasses $b_2 \ldots b_k$ at the roots of function (13) are regarded as nuisance parameters, which need not necessarily have any obvious relationship to the confidence intervals for the biomass estimates within

each stratum. The values of p_j and κ_j^2 are also nuisance parameters which have to be found by maximising the log-likelihood function for each trial solution $B,b_2...b_n$. Thus, numerical maximisations occur in a nested fashion, which results in a considerable computational burden.

A computer program (TRAWLCI) implementing this method has been developed and made available to CCAMLR.

SIMULATION TESTS OF THE METHOD

The computer code implementing the method was embedded in a computer program which generated random data from a number of strata, for which the true densities and biomasses were known. A large number of such data sets was generated and the method used to calculate biomass estimates and their 95% confidence intervals so as to compare them with the true parameters used to generate the data.

Three strata were used, with the following parameters:

Stratum	Area	Biomass	Mean Density	CV	P
1	1 000	10 000	10	5	0.1
2	2 000	10 000	5	2	0.5
3	10 000	10 000	1	1	0.9

All the possible combinations of these three strata were used in the trials. The results are shown in Table 1. Fifty hauls were made in each stratum, and so the expected numbers of non-zero data in each stratum are 5, 25 and 45 respectively; the actual number of non-zero hauls is a binominal random variable. The results show that the abundance estimates appear to be unbiased. The results from stratum 1 do not seem to be very close to the true value, but the differences are not statistically significant given the number of trials conducted. The apparent differences are therefore probably due to the estimates having a large coefficient of variation (CV), and a small number of non-zero data points. The number of non-zero data points has a large effect on the precision of the estimates of the parameters of the lognormal component of the delta distribution.

The lower confidence bound estimates appear to be quite reliable and close to the nominal percentage (2.5%). The upper confidence bounds appear to be slightly too close to the mean, with worst performance in the cases which include stratum 1. Overall, the confidence intervals are close to the nominal percentage of coverage, with the worst departures occurring in cases where one of the strata has a high CV and few non-zero hauls.

CONCLUDING REMARKS

Straightforward extensions using the log-likelihood function enable likelihood ratio tests of hypotheses such as whether the density estimates from different surveys are homogeneous, or have specific hypothesised values. They also provide the basis for fitting regressions to time series of trawl survey abundance estimates, or mixture distributions to density-at-length data. The latter procedure has been used in estimating proportions of recruits in krill surveys (de la Mare, 1984).

ACKNOWLEDGEMENTS

I am grateful to Doug Butterworth and an anonymous reviewer for helpful comments and for noticing some notational inconsistencies in the first draft of this paper.

REFERENCES

Aitchison, J. 1955. On the distribution of a positive random variable having a discrete probability mass at the origin. *J. Am. Stat. Assoc.*, 50: 901-908.

Cox, D.R. and D.V. Hinkley. 1974. *Theoretical Statistics*. Chapman and Hall, London.

de la Mare, W.K. 1994. Estimating krill recruitment and its variability. *CCAMLR Science* (this volume).

Pennington, M. 1983. Efficient estimators of abundance for fish and plankton surveys. *Biometrics*, 39: 281-286.

Saville, A. (Ed.). 1977. Survey methods of appraising fishery resources. *FAO Fish. Tech. Pap.*, 171: 76 pp.

SC-CAMLR. 1992. CCAMLR Workshop on Design of Bottom Trawl Surveys. In: *Report of the Eleventh Meeting of the Scientific Committee (SC-CAMLR-XI)*, Annex 5, Appendix G. CCAMLR, Hobart, Australia: 289-329.

Table 1: Results of simulation tests of the trawl survey confidence interval (C.I.) estimator. In cases involving stratum 1, the number of trials is affected by the exclusion of data sets which had less than two non-zero observations in that stratum. The exclusion of data sets with less than two non-zero observations does not affect the conclusions as to whether the estimates are unbiased or about the properties of the confidence intervals, which can only be calculated given this condition. The mean abundance column is the mean over the complete set of trials for that combination of strata. The column Low C.I. > True gives the percentage of trials for which the estimated lower confidence interval for abundance is greater than the true value. The column Upper C.I. < True is defined in an analogous way. The final column is the total percentage of trials in which the estimated confidence interval does not include the true abundance.

Stratum Combination	Number of Trials	True Abundance	Mean Abundance	Standard Error	Low C.I. > True (%)	Upper C.I. < True (%)	C.I. Misses True %
1	9 660	10 000	9 929	70.7	2.03	6.40	8.43
2	10 000	10 000	10 023	27.6	2.20	3.61	5.81
3	10 000	10 000	9 999	13.8	2.37	3.42	5.79
1+2	966	20 000	19 409	235.5	2.4	3.7	6.1
1+3	965	20 000	19 660	244.3	2.6	7.5	10.1
2+3	1 000	20 000	20 032	97.5	2.5	2.8	5.3
1+2+3	964	30 000	29 427	247.7	2.6	4.5	7.1

Légendes des tableaux

Tableau 1: Résultats des tests par simulation du paramètre d'estimation de l'intervalle de confiance (C.I.) de la campagne d'évaluation par chalutages. Lorsque la strate 1 est concernée, le nombre d'essais est affecté par l'exclusion des jeux de données comprenant moins de deux observations non nulles dans cette strate. L'exclusion des jeux de données comprenant moins de deux observations non nulles n'affecte pas les conclusions déterminant si les estimations sont biaisées ou non ou concernant les propriétés des intervalles de confiance qui ne peuvent être calculés que si cette condition est remplie. La colonne d'abondance moyenne (mean abundance) est la moyenne calculée sur une série entière d'essais pour ces strates combinées. La colonne "Low C.I. > True" donne le pourcentage d'essais pour lesquels la limite inférieure de l'intervalle de confiance estimé pour l'abondance est supérieure à la valeur réelle. La colonne "Upper C.I. < True" est définie d'une manière analogue. La dernière colonne est le pourcentage total d'essais dans lesquels l'abondance réelle n'est pas comprise dans l'intervalle de confiance estimé.

Список таблиц

Таблица 1: Результаты имитационного испытания определителя доверительного интервала (C.I.) по траловым съемкам. В случае слоя 1 исключение наборов данных, имеющих менее двух "не-нулевых" наблюдений, влияет на количество испытаний. Исключение наборов данных, имеющих менее двух "не-нулевых" наблюдений не влияет на выводы о том, являются ли оценки несмещенными или заключения о характеристиках доверительных интервалов, поддающихся вычислению только при этом условии. Колонка "средняя численность" показывает среднюю величину для всех испытаний по данной комбинации слоев. олонка "Low C.I. > True" показывает процентное отношение испытаний, для которых оцененный нижний доверительный интервал больше истинного значения. Колонка "Upper C.I. < True" определен аналогичным образом. Последняя колонка - общее процентное отношение испытаний, в которых оцененный доверительный интервал не включает истинную численность.

Lista de las tablas

Tabla 1: Resultados de las pasadas de simulación del estimador del intervalo de confianza de la prospección de arrastre (C.I.). En los casos que incluyen el estrato 1, el número de pruebas se ve afectado por la omisión de los grupos de datos que tienen menos de dos observaciones no nulas en ese estrato. La omisión de grupos de datos con menos de dos observaciones no nulas no afecta la conclusión de que los cálculos son correctos o acerca de las propiedades de los intervalos de confianza. Estos intervalos pueden calcularse sólo si se cumple esta condición. La columna de abundancia promedio es la media de todas las pasadas para esa combinación de estratos. La columna C.I. Bajo > Real proporciona el porcentaje de pasadas para el cual el calculado intervalo inferior de confianza de abundancia es mayor que el valor verdadero. La columna C.I. Superior < Real se define de modo análogo. La columna final representa el porcentaje total de las pasadas en la que el intervalo de confianza calculado no incluye la abundancia real.