Analysing differences between two age determination methods by tests of symmetry

J.M. Hoenig, M.J. Morgan, and C.A. Brown

Abstract: A common problem in fisheries science is the comparison of two methods for obtaining ages of individual animals. Often, indices of precision are computed for this purpose. We believe such indices are inappropriate both as measures of precision and for comparative purposes because they do not properly account for age effects or take the experimental design into consideration. We suggest that if the overall level of agreement is low between two ageing methods used on the same sample of fish, then one can use a test of symmetry to look for evidence of systematic disagreement. A χ^2 test is used to determine if the number of fish assigned age i from method 1 and age j from method 2 differs significantly from the number of fish assigned age j from method 1 and age i from method 2. Such a test can also be used to determine the range of nominal ages over which two methods appear to give comparable results.

Résumé : Dans le domaine de l'halieutique se pose souvent le problème de la comparaison de deux méthodes pour obtenir l'âge d'un animal. On calcule souvent à cette fin des indices de précision. Nous pensons que ces indices sont inappropriés tant comme mesure de la précision que comme outils de comparaison parce qu'ils ne rendent pas compte correctement des effets de l'âge et ne tiennent pas compte du protocole expérimental. Nous posons que, si le niveau global de concordance est faible entre deux méthodes de détermination de l'âge employées sur un même échantillon de poisson, on peut avoir recours à un test de symétrie pour rechercher les preuves de divergence systématique. On se sert d'un test χ^2 pour déterminer si le nombre de poissons auxquels la méthode 1 assigne l'âge i et la méthode 2 l'âge j diffère de façon significative par rapport au nombre de poissons auxquels la méthode 1 assigne l'âge j et la méthode 2 l'âge i. Un tel test peut aussi servir à déterminer la plage d'âge nominal dans laquelle les deux méthodes semblent donner des résultats comparables.

[Traduit par la Rédaction]

Introduction

Common methods for evaluating precision or consistency among age determinations are that of Beamish and Fournier (1981) and the modification of their method by Chang (1982). These methods calculate what are called indices of precision for a series of age determinations. A set of determinations with a small index of precision is considered precise. Both Beamish and Fournier (1981) and Chang (1982) suggested that indices of precision are appropriate for comparing two ageing methods or two age readers.

Beamish and Fournier (1981) calculated average percent error (APE) as

$$[1] \quad \mathrm{APE_{BF}} = 100 \times \frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{R}\sum_{i=1}^{R}\frac{|X_{ij} - X_j|}{X_j}\right)$$

where N = the number of fish aged, R = the number of times each fish is aged, X_{ij} = the ith age determination for the jth fish, and X_j = the average age calculated for the jth fish.

Chang (1982) calculated APE using the square root of the sum of squared deviations instead of the sum of absolute values:

$$[2] \quad \mathrm{APE_C} = 100 \times \frac{1}{N}\sum_{j=1}^{N}\frac{\sqrt{\sum_{i=1}^{R}\frac{(X_{ij} - X_j)^2}{R-1}}}{X_j}.$$

These methods are based on the assumption that the occurrence of several age groups in the sample is accounted for by expressing the precision in relative terms. It is assumed that in this way the index of precision does not depend on the age of fish and one overall index can be calculated from all of the age groups. However, there may still be differences in precision among ages which these procedures obscure. Brown (1988) calculated an average

Received March 7, 1994. Accepted July 22, 1994.
J12302

J.M. Hoenig and M.J. Morgan. Department of Fisheries and Oceans, P.O. Box 5667, St. John's, NF A1C 5X1, Canada.
C.A. Brown. Rosenstiel School of Marine and Atmospheric Sciences, University of Miami, 4600 Rickenbacker Causeway, Miami, FL 33149, USA.

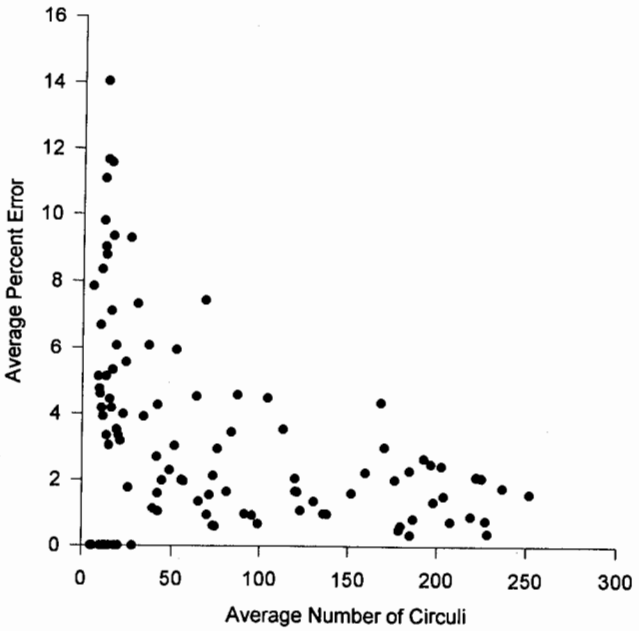
percent error for each lemon shark (*Negaprion brevirostris*) examined according to the method of Beamish and Fournier (1981) and then plotted these indices against the average age assigned. Ages were determined by counting growth marks in vertebral centra. He found that the APE as well as the variability in APE was much greater for younger lemon sharks than for older ones (Fig. 1).

Kimura and Lyons (1991) concluded that relative precision (equation [2] and minor variations thereof) was independent of age for six species examined and, therefore, that it was appropriate to average an index of precision over fish of different ages. However, if their indices are plotted against age, there are clear trends with age for three of the species they studied (walleye pollock (*Theragra chalcogramma*), Atka mackerel (*Pleurogrammus monopterygius*), and sablefish (*Anaplopoma fimbria*)). Furthermore, the indices were zero for the youngest age for all six species, suggesting that in general the youngest fish are aged with greater relative precision than are the older fish. The possibility of finding trends with age has also been pointed out by Campana and Jones (1992). In a comparison of two age readers, reading different samples, if the two samples contain different proportions of small fish, there will be apparent differences between the readers. This is because there is greater relative precision in ageing young fish than old fish, even after supposedly accounting for age. Thus, indices of precision cannot be averaged across age groups.

The methods of Beamish and Fournier (1981) and Chang (1982) also assume that the variability among observations of individual fish can be averaged over all fish. However, variability in age determinations can come from a variety of sources (i.e., within fish, among fish, among ages, etc.). If the components of the overall variance are not separated, it becomes difficult to interpret the index and to compare index values from different studies that had different experimental designs. An APE from reading two fish 10 times is not comparable with one from reading 10 fish two times. In the first, most of the variability would be within fish whereas in the second, it would be mainly among fish. To account for the different sources of variability properly, one would need to do an analysis of variance and variance components (Kimura et al. 1979). However, because the response variable is observed age, it is inappropriate to do separate analyses for each "nominal" age group as in Kimura et al. (1979). We suggest that separate analyses be conducted by size group (an independent variable correlated with age). This will help ensure that the assumption of homogeneous variances is met.

Finally, we join Campana and Jones (1992) in pointing out that comparisons of precision are only of interest if there is no evidence of systematic disagreement among readers or methods. Unfortunately, this appears to be frequently overlooked. For example, Beamish and Fournier (1981) and Chang (1982) described a study in which two readers aged the same 20 fish three times. They pointed out that reader No. 2 in their example appeared to be less precise than reader No. 1. What this meant in practice was that, for most (13 out of 20) fish, the oldest age assigned differed from the lowest age by 1 year for reader No. 2. In contrast, for most (11 out of 20) fish, the oldest age was

Fig. 1. Plot of APE against average number of circuli in vertebral centra of individual lemon sharks. The number of circuli was counted three times for each shark. Circuli are formed with a lunar periodicity (Brown 1988).



equal to the youngest age for reader No. 1. All other things being equal, we would conclude that reader No. 1 was more precise. However, neither Beamish and Fournier (1981) nor Chang (1982) mentioned the fact that 53% of the time the observation of reader No. 1 was above that of reader No. 2 whereas 8% of the time reader No. 1 was below reader No. 2 (and 38% of the time they agreed). Thus, there is clear evidence that reader No. 1 tended to assign older ages than reader No. 2. The question of which (if either) reader was unbiased seems more important than who was more precise. Furthermore, it is apparent that the first reading for each fish made by reader No. 2 tended to be the highest made by that reader whereas the second reading tended to be lowest. Reader No. 2 appears to have changed criteria over time and there is no way to predict how this reader would perform at other times.

It would seem that these indices tend to oversummarize the data and are hard to interpret. They may lead to false conclusions when used to compare the precision of different readers or methods when the studies compared are based on different experimental designs. The questions that can be addressed by examining measures of precision are not relevant if there are systematic differences among readers or methods, or if performance changes over time. We suggest that a more appropriate approach to comparing age determinations from different structures or readers is a test of symmetry.

In this paper, we describe the test of symmetry developed by Bowker (1948) and show how it can be used to compare two different ageing methods which use either different structures or different readers to determine age. We also show how the test can be extended to determine where differences between methods occur.

A test of symmetry

Valid comparisons of methods can only be made if the methods are used on the same statistical population. Statistical efficiency is attained if the two methods are used on the same individual fish. We assume here that the same fish are aged by two readers or by two methods.

The first step in a comparison of ageing methods would be to determine how often the two methods agree. If agreement is high, there may be no need to proceed further. If, however, there is a considerable amount of disagreement between methods, it is important to know whether this disagreement is simple random error of if there is a systematic difference between methods.

Bowker's (1948) method was designed to test the hypothesis that an $m \times m$ contingency table consisting of two classifications of a sample into categories (e.g., ages given by two readers) is symmetric about the main diagonal. The test statistic is distributed as a chi-square variable with $m(m-1)/2$ degrees of freedom for a table that has no empty cells. The test statistic is

$$[3] \quad X^2 = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}$$

where n_{ij} = the observed frequency in the ith row and jth column and n_{ji} = the observed frequency in the jth row and ith column.

The summation is over all the cells above the diagonal. These cells are paired with the corresponding cells below the diagonal. If there is a systematic difference between methods, then the test statistic will tend to be large. If, however, the differences are due to simple random error, then the value of n_{ji} will be very similar to that of n_{ij} (the subscripts/readers being essentially interchangeable) and the test statistic will not be significant. The number of degrees of freedom is equal to the number of comparisons. If both cells in a pair (n_{ji} and n_{ij}) are zero, the pair is dropped from the test statistic and the degrees of freedom is reduced by 1.

Example

We consider data from O'Gorman et al. (1987) in which a comparison is made between age determined by otoliths and age determined by scales for alewife (*Alosa pseudoharengus*) from Lake Huron (Table 1A). In O'Gorman et al. (1987), cells with no observations were left blank. To illustrate the pairing of the data, we have explicitly added zeros to those cells needed for calculating the test statistic. Thus, the bottom four rows do not appear in the table presented in O'Gorman et al. (1987).

In Table 1A the number in a cell, n_{ij}, is the number of times a fish was assigned to age j using an otolith when it was assigned to age i using a scale. The numbers are printed in bold typeface in the cells where the two methods agree. The two methods agree only 58% of the time. Differences between the methods should thus be investigated. It is clear that the scale method does not produce ages older than 6 whereas the otolith method yields ages as old as 10. This indicates that a test of symmetry will probably be significant. The test statistic is

$$\begin{aligned} X^2 = {} & \frac{(1-0)^2}{1+0} + \frac{(0-3)^2}{0+3} + \frac{(2-4)^2}{2+4} + \frac{(5-0)^2}{5+0} \\ & + \frac{(4-1)^2}{4+1} + \frac{(2-0)^2}{2+0} + \frac{(1-0)^2}{1+0} + \frac{(2-0)^2}{2+0} \\ & + \frac{(3-0)^2}{3+0} + \frac{(5-0)^2}{5+0} + \frac{(1-0)^2}{1+0} + \frac{(1-0)^2}{1+0} \\ & + \frac{(4-0)^2}{4+0} + \frac{(1-0)^2}{1+0} + \frac{(1-0)^2}{1+0} + \frac{(2-0)^2}{2+0} \\ = {} & 34.47, \text{ df} = 16, p = 0.005. \end{aligned}$$

Here the number of degrees of freedom (16) is equal to the number of comparisons made because the table has pairs of cells which are empty. As expected, the hypothesis of symmetry is rejected presumably because older ages do not occur in the scale readings.

An extension to the test

To determine at which point the two methods begin to differ, the table can be collapsed to form a "plus" group. We compared the two methods with fish greater than or equal to age 4 placed in the ≥4 plus group. To do this, any cell with both subscripts greater than or equal to 4 is added to produce the ≥4 plus group (Tables 1B and 1C). All of the remaining cells with the row subscript greater than 4 are added to the fourth row cell with the same column subscript (e.g., $n_{5,3}$ and $n_{6,3}$ are added to $n_{4,3}$). The same procedure is used for cells with the column subscript greater than 4 (e.g., $n_{3,5}$ and $n_{3,6}$ are added to $n_{3,4}$). In this case, the test statistic is $X^2 = 12.67$, df = 4, $p = 0.013$, and again we reject the hypothesis of symmetry.

This process can be taken a step further, to compare the two methods with a ≥3 plus group (box in Table 1C). Here the test statistic would be $X^2 = 4.7$, df = 3, $p = 0.195$. The hypothesis of symmetry cannot be rejected and we have no significant evidence that scales and otoliths would produce different estimates of age composition if ages are assigned to 0, 1, 2, or ≥3. (Note that the test is based on only 10 fish and 3 degrees of freedom, so it does not have much statistical power.)

Discussion

According to various authors, the percentage agreement statistic is inappropriate and should not be calculated (Beamish and Fournier 1981; Chang 1982). We disagree and find that this statistic is intuitive and important for decision making. If agreement is high, there may be no point in investigating further because the impact of the disagreements on the estimated age composition is minimal. We note, however, that the level of agreement is likely to vary with the age composition of the sample so that the percentage agreement is not interpretable as a property of the species, stock, reader, or method.

Indices of precision can be calculated for individual fish. However, their interpretation becomes problematic if the values are averaged over fish (even if the fish are all of the same nominal age). This is because within-fish and

Table 1. (A) Ages of alewife from Lake Huron as determined from otoliths and scales (from O'Gorman et al. 1987). Bold numbers are where the two methods agree. Numbers with the same superscript are compared in the test of symmetry. In this example, there were no observations below the solid line, which makes asymmetry easily detectable by visual examination of the table. (B) Formation of a table with a ≥4 plus group from the complete table. The UPPER area remains the same when the table is collapsed; in the LOWER area the cells have both subscripts ≥4 and all of these are added to cell 4,4. Cells 5,3 and 6,3 have row subscripts >4 and are added to cell 4,3. Cells 3,5 and 3,6 have column subscripts >4 and are added to cell 3,4. (C) Resulting collapsed table with the ≥4 plus group; also indicated is the way to form a ≥3 plus group.

(A)

Scale age \ Otolith age	0	1	2	3	4	5	6	7	8	9	10
0	**2**	$1^{(1)}$									
1	$0^{(1)}$	**13**	$0^{(2)}$								
2		$3^{(2)}$	**16**	$2^{(3)}$							
3			$4^{(3)}$	**11**	$5^{(4)}$	$2^{(8)}$	$1^{(11)}$				
4				$0^{(4)}$	**12**	$4^{(5)}$	$3^{(9)}$	$1^{(12)}$	$1^{(15)}$		
5				$0^{(8)}$	$1^{(5)}$	**4**	$2^{(6)}$	$5^{(10)}$	$4^{(13)}$		
6				$0^{(11)}$	$0^{(9)}$	$0^{(6)}$	**2**	$1^{(7)}$		$1^{(14)}$	$2^{(16)}$
7					$0^{(12)}$	$0^{(10)}$	$0^{(7)}$				
8					$0^{(15)}$	$0^{(13)}$					
9							$0^{(14)}$				
10							$0^{(16)}$				

(B)

Scale age \ Otolith age	0	1	2	3	4	5	6	7	8	9	10
0	**2**	$1^{(1)}$									
1	$0^{(1)}$	**13**	$0^{(2)}$								
2		$3^{(2)}$	**16**	$2^{(3)}$							
3	UPPER		$4^{(3)}$	**11**	$5^{(4)}$	$\leftarrow 2^{(8)}$	$\leftarrow 1^{(11)}$				
4				$0^{(4)}$	**12**	$4^{(5)}$	$3^{(9)}$	$1^{(12)}$	$1^{(15)}$		
5				$\uparrow$ $0^{(8)}$	$1^{(5)}$	**4**	$2^{(6)}$	$5^{(10)}$	$4^{(13)}$		
6				$\uparrow$ $0^{(11)}$	$0^{(9)}$	$0^{(6)}$	**2**	$1^{(7)}$		$1^{(14)}$	$2^{(16)}$
7					$0^{(12)}$	$0^{(10)}$	$0^{(7)}$				
8					$0^{(15)}$	$0^{(13)}$					
9							$0^{(14)}$			LOWER	
10							$0^{(16)}$				

(C)

Scale age \ Otolith age	0	1	2	3	≥4
0	**2**	$1^{(1)}$			
1	$0^{(1)}$	**13**	$0^{(2)}$		
2		$3^{(2)}$	**16**	$2^{(3)}$	
3			$4^{(3)}$	**11**	$8^{(4)}$
≥4				$0^{(4)}$	**43**

Table 2. Hypothetical situation in which asymmetry will be detected by Bowker's χ^2 test but not by two other tests. The marginal row frequencies equal the marginal column frequencies so the hypothesis that the two marginal distributions are equal will not be rejected. The number of observations above the diagonal (40) equals the number of observations below the diagonal so the hypothesis of equal proportions above and below the diagonal will not be rejected by a sign test. Superscripts indicate cells which are paired for Bowker's test.

		Reader A				
	Age	1	2	3	4	Marginal frequency
Reader B	1	**30**	$0^{(1)}$	$20^{(2)}$		50
	2	$20^{(1)}$	**30**		$0^{(3)}$	50
	3	$0^{(2)}$		**30**	$20^{(4)}$	50
	4		$20^{(3)}$	$0^{(4)}$	**30**	50
Marginal frequency		50	50	50	50	

between-fish variability are not distinguished and age effects are ignored.

Our main concern with the use of indices of precision is that they can draw attention away from important trends in the data. For example, in three articles proposing methodology for comparative studies of ageing programs, indices of precision were calculated but the trends in the data across ages and readers were not discussed (Beamish and Fournier 1981; Chang 1982; Kimura and Lyons 1991). Furthermore, indices of precision are often not comparable. Within a species, APEs among samples can differ solely because age compositions differ; the precision of a method is likely to change if readers change, and studies using different designs may have different APEs because of the differences in design rather than inherent differences in precision. Thus, a value of APE reported in the literature is generally not interpretable.

In this paper, we advocate the examination of two-way contingency tables for asymmetry. There are a variety of statistical tests that can be used to test for asymmetry such as a sign test (H_0: proportion of observations above the diagonal equals the proportion below), a test of marginal homogeneity (H_0: age composition according to method 1 equals age composition according to method 2), and the Bowker-type χ^2 test. Some tests are extremely powerful for detecting specific types of asymmetry at the expense of the ability to detect other types of asymmetry. The Bowker-type test, however, is the most general in the sense that it looks at all parts of the table off the diagonal and can detect asymmetries that are not detectable by other methods (Table 2). Additionally, the contribution of different cells of the table to the overall χ^2 test statistic can be examined to determine where differences arise.

Acknowledgements

We thank G. Evans, W. Warren, A.R. Kronlund, and an anonymous reviewer for helpful comments.

References

Beamish, R.J., and Fournier, D.A. 1981. A method for comparing the precision of a set of age determinations. Can. J. Fish. Aquat. Sci. **38**: 982–983.

Bowker, A.H. 1948. A test for symmetry in contingency tables. J. Am. Stat. Assoc. **43**: 572–574.

Brown, C.A. 1988. Validated age assessment of the lemon shark, *Negaprion brevirostris*, using tetracycline labeled vertebral centra. Thesis, University of Miami, Miami, Fla.

Campana, S.E., and Jones, C.M. 1992. Analysis of otolith microstructure data. *In* Otolith microstructure examination and analysis. *Edited by* D.K. Stevenson and S.E. Campana. Can. Spec. Publ. Fish. Aquat. Sci. **117**: 73–100.

Chang, W.Y.B. 1982. A statistical method for evaluating the reproducibility of age determination. Can. J. Fish. Aquat. Sci. **39**: 1208–1210.

Kimura, D.K., and Lyons, J.J. 1991. Between-reader bias and variability in the age-determination process. Fish. Bull. U.S. **89**: 53–60.

Kimura, D.K., Mandapat, R.R., and Oxford, S.L. 1979. Method, validity, and variability in the age determination of yellowtail rockfish (*Sebastes flavidus*), using otoliths. J. Fish. Res. Board Can. **36**: 377–383.

O'Gorman, R., Barwick, D.H., and Bowen, C.A. 1987. Discrepancies between ages determined from scales and otoliths for alewives from the Great Lakes. *In* Age and growth of fish. *Edited by* R.C. Summerfelt and G.E. Hall. The Iowa State University Press, Ames, Iowa. pp. 203–210.