International Commission for the Northwest Atlantic Fisheries

Serial No. 3173
(D.c.9)

ICNAF Res. Doc. 74/26

ANNUAL MEETING - JUNE 1974

Bias in Two Length Frequency Formulae[1]

by

W. G. Doubleday
Department of the Enviroment
Fisheries and Marine Service
Biological Station
St. Andrews, N.B.

Abstract

Two formulae currently used to estimate the frequency distribution of lengths of groundfish in Canadian commercial landings are examined. The nature of the bias of each formula is explained and ways are suggested to reduce the biases.

Introduction

Canadian commercial groundfish sampling is carried out on a portion of the landings at Canadian ports. The length frequency distribution of fish in unsampled landings is assumed to be the same as that of fish in sampled landings. Discards at sea are ignored. In what follows, attention is directed at the analysis, not the collection, of sampling data. Therefore, biases due to current sampling methods are ignored.

Usually, fish are landed either as one category (ungraded) or as three categories (small, medium, and large). If the landing being sampled has three categories, then a separate sample is taken from each category. Ordinarily, the fish are sampled from boxes and a sample consists of all the fish in several boxes. It is assumed in the following analysis that each fish in a market category in a landing is equally likely to appear in the sample of that category from that landing. This is a weaker assumption than to assume simple random sampling within a category from a landing.

The difficulty in the statistical analysis is due to the unknown number of fish landed in each market category. This number is estimated for each landing from the total weight of fish landed in that category and either the weight of the sample or a length-weight key determined from research vessel cruises. The length-weight key currently in use (Kohler et al., MS 1970) is a linear regression of log (length) on log (weight). Biases in the length-weight key and in the determination of the sample weight are here ignored, although it is worth mentioning that inaccuracies in the determination of sample weights due to variations in the weight of boxes prompted the adoption of the length-weight key.

Biases due to variations in the stock being fished are minimized by restricting estimates to three-month periods in limited areas. These effects are also ignored.

[1] Presented to the Special Commission Meeting, FAO, Rome, January 1974, As Res. Doc. 74/26

Two Formulae

The following notation is adopted. Landings sampled are symbolized by i, i=1,2, ..., I ; market categories by j, j=1 for large fish, j=2 for medium fish, and j=3 for small fish; and length classes by k, k=1,2, ..., K.

Wij = total weight of class j in landing i
wij = weight of sample from class j in landing i
Xijk = number of fish in class k from category j in landing i
LW_k = weight assigned to a fish from length class k by the length-weight key.

All weights are in pounds and all lengths in cm.

The following formula was suggested by Messrs. Pinhorn and Sandeman for estimation of the number of fish in length class k in category j in landing i. (They considered only one category since their fish were not sorted into size categories.)

$\frac{Wij}{wij}$ Xijk = estimated number of fish landed in class k in category j in landing i

Then the total number landed in length class k in category j is estimated by

$$\sum_{i=1}^{I} \frac{Wij}{wij} Xijk \qquad (1)$$

This estimate leads to the following estimate for the % frequency of length class k in category j.

$$100 \times \sum_{i=1}^{I} \frac{Wij}{wij} Xijk \Bigg/ \sum_{k=1}^{K}\left(\sum_{i=1}^{I} \frac{Wij}{wij} Xijk\right) \qquad (2)$$

It is desirable to extend Pinhorn and Sandeman's formula so that comparisons can be made with the formula in use at St. Andrews. The following estimate of the % frequency of class k for all landings is proposed

$$100 \times \sum_{i=1}^{I}\left(\sum_{j=1}^{3} \frac{Wij}{wij} Xijk\right) \sum_{j=1}^{3}\left(\sum_{k=1}^{K}\left(\sum_{i=1}^{I} \frac{Wij}{wij} Xijk\right)\right) \qquad (3)$$

The estimation procedure in use at St. Andrews has two stages. First, the weights wij of the various samples are estimated using the length-weight key.

$$\bar{w}ij = \sum_{k=1}^{K} Xijk \ LW_k \qquad (4)$$

Then the % length frequencies are estimated by pooling all samples as follows:

$$100 \times \sum_{j=1}^{3}\left\{\frac{\left(\sum_{i=1}^{I} Xijk\right)\left(\sum_{i=1}^{I} Wij\right)}{\left(\sum_{i=1}^{I} \bar{w}ij\right)\left(\sum_{j=1}^{3}\left(\frac{\sum_{i=1}^{I} Wij}{\sum_{i=1}^{I} \bar{w}ij}\left(\sum_{i=1}^{I}\sum_{k=1}^{K} Xijk\right)\right)\right)}\right\} \qquad (5)$$

In this formula, all samples from the same category are pooled to form one large sample, the weights of the landings for each category are combined before dividing by the sum of calculated sample weights to give an estimate of the total number of fish in that category in all sampled landings.

Example

Table 1 contains samples from two cod landings in the fourth quarter of 1949 from the Banquereau ground off eastern Nova Scotia. The appropriate length-weight key (Kohler et al., MS 1970, Table II) is included. The formulae are illustrated by the calculation of the % frequencies for classes 58 cm, 82 cm, and 100 cm.

For the 58 cm class, the extended Pinhorn and Sandeman formula gives an overall % frequency of

$$100 \left\{\frac{19309 \times 76 + 23376 \times 28 + 4837 \times 2}{425\times20+2287\times95+19309\times120+22376\times280+4837\times85}\right\} = 22.44$$

For the 82 cm class, the estimate is:

$$100 \left\{\frac{19309 \times 1 \quad + 2287 \times 1}{425\times20+2287\times95+19309\times120+22376\times280+4837\times85}\right\} = 0.227$$

For the 100 cm class, the estimate is:

$$100 \left\{\frac{425 \times 1 \quad + 2287 \times 14}{425\times20+2287\times95+19309\times120+22376\times280+4837\times85}\right\} = 0.342$$

For the St. Andrews formula,

$\overline{W}11 = 370.05$, $\overline{W}12 = 939.35$, $\overline{W}13 = 0$

$\overline{W}21 = 1422.93$, $\overline{W}22 = 424.01$, $\overline{W}23 = 198.78$

For the 58 cm class, the estimated % frequency is:

$$100 \times \left\{\frac{104 \times (60340 + 77920)}{(1939.35 + 424.01) \times 48382} + \frac{2 \times (11380)}{198.78 \times 48382}\right\} = 22.04$$

For the 82 cm class, the estimate is:

$$100 \times \left\{\frac{1 \times (7430 + 38520)}{(370.05 + 1422.93) \times 48382} + \frac{1 \times (60340 + 77920)}{(939.35 + 424.01) \times 48382}\right\} = 0.263$$

For the 100 cm class, the estimate is:

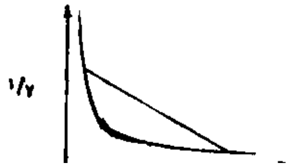
$$100 \times \left\{\frac{15 \times (7430 + 38520)}{(370.05 + 1422.93) \times 48382}\right\} = 0.795$$

Ratio Estimates

In the Pinhorn and Sandeman formula, the quantity $\sum_{i=1}^{I} \frac{Wij}{wij} Xijk$ is used to estimate the total number landed in class j in the sampled landings. The terms of the sum, i.e. $\frac{Wij\ Xijk}{wij}$ are estimates of the number of fish landed in class j in the i'th landing. The quantity $\frac{Xijk}{wij}$ is the ratio of two random variables, unless the sample weight is constant. (As was remarked earlier, the nominal weight of a box of fish can differ considerably from its actual weight.) Ordinarily the expected ratio of two random variables, X and Y, is not equal to the ratio of their individual expected values

$$E(X/Y) \neq E(X)/E(Y)$$

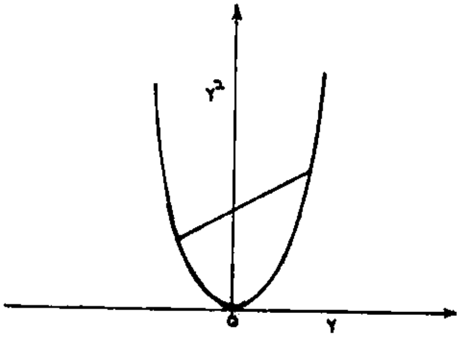
First, consider the random variable Y. 1/Y is a transformation of Y. If the reciprocal of a number, 1/y, is plotted against the number, y, the following graph results:



E 4

26

If a line is drawn joining any two points on the graph (a secant) the part of the graph between the end points of the line lies below the line. Graphs like this, and the corresponding transformations are called convex. A more familiar convex transformation is the one which plots the square of a number, y^2, against the number, y.



If probabilities are given to the possible values y, then the random variable Y is transformed to Y^2. Now, $E(Y^2) - (E(Y))^2$ is the variance of Y, and it is therefore greater than or equal to zero. Therefore

$$E(Y^2) \geqslant (E(Y))^2 \qquad (6)$$

and $E(Y^2) = (E(Y))^2$ only if Y takes on only one value with probability one.

The inequality (6) is true for any convex transformation of Y. In general, E(transformed Y) $\geqslant$ transformed E(Y) with equality only if Y is limited to one value. This property is called Jensen's inequality. If the inequality is applied to the reciprocal transformation, then

$$E(1/Y) \geqslant 1/E(Y)$$

If X and Y are statistically independent, then so are X and 1/Y, and hence $E(X/Y) = E(X)E(1/Y) \geqslant E(X)/E(Y)$.

If X and Y are dependent, the situation is more complicated, but there is the following approximation for large samples (Hansen et al., 1952, p 112):

$$E(X/Y) - E(X)/E(Y) \approx \frac{E(x)}{E(y)} \left((CV(Y))^2 - \rho xy CV(X) CV(Y) \right) \qquad (7)$$

The bias of the ratio estimate is small when the coefficient of variation (CV) of the denominator is small. Thus, if large samples are taken within a ship, the bias is reduced since the coefficient of variation of the weight of the sample is inversely proportional to the square root of the sample size.

Unfortunately, combining the estimates for the various ships in the sample by a weighted average also combines the biases by the same weighted average, so that this source of bias is not reduced by sampling a large number of ships.

The correlation ρ between the number of fish in a given length category with the total weight of the sample in samples containing a fixed number of fish is negative for classes of short length and positive for classes of large length. Thus, the bias is different for the different length classes.

Weighting

Since samples from different landings may be considered statistically independent, the pooling of samples within a market category reduces the bias due to the ratio estimation of numbers landed. This is because the coefficient of variation of the average of the sample weights tends to zero as the number of landings increases. However, the St. Andrews formula has another source of bias.

E 5

In the St. Andrews formula, the weight given to a sample from a landing is proportional to the size of the sample, not necessarily to the size of the landing of the appropriate market category. In the example, category 2 is sampled three times as heavily on the first landing as on the second (1.5% vs 0.5%). These departures from proportional sampling are unimportant unless large catches have a different length composition than small catches.

Suppose that the frequency of a particular class is M in the total landings of a given category and Mi in the i'th landing. Also suppose that the fraction of the total landings of that category which is made up of fish from the i'th landing is Pi. Let the weights assigned by the St. Andrews formula be Pi. Then the following relations hold:

$$M = \sum_{i=1}^{I} P_i M_i = \sum_{i=1}^{I} P_i' M \tag{8}$$

$$M - \sum_{i=1}^{I} P_i' M_i = \text{bias due to } P_i' \text{ differing from } P_i$$

$$= \sum_{i=1}^{I} P_i' M - \sum_{i=1}^{I} P_i' M_i$$

$$= \sum_{i=1}^{I} P_i' (M - M_i) \tag{9}$$

This quantity (9) is nonzero if the Mi of overrepresented landings are consistently greater than M or consistently less than M.

Suggestions

The formula (7) can be rewritten as follows:

$$E\left(\frac{X}{Y}\right) \approx \frac{E(X)}{E(Y)} \left\{ 1 + \left((CV(Y))^2 - \rho xy CV(X) CV(Y) \right) \right\} \tag{10}$$

The terms in brace brackets in (10) can be estimated from accumulated commercial sampling data to give a correcting factor for the terms

$$\frac{W_{ij} \quad X_{ijk}}{w_{ij}}$$

in Pinhorn and Sandeman's formula. The correcting factor will depend on the sample size.

The bias due to incorrect weighting in the St. Andrews formula can be removed by making the sample size within a category and landing proportional to the number of fish in that landing and category. A good approximation would be obtained by taking a systematic sample of every twentieth box of category 1, every one hundredth box of category 2, and every sixtieth box of category 3 in each landing. The actual box to be sampled can be determined by choosing a number from one to twenty from a table of random numbers.

Acknowledgements

Thanks are due to Mr. A. T. Pinhorn for bringing attention to this problem, and to Mr. D. N. Fitzgerald for providing the data used in the example.

References

Hansen, M. H., W. N. Hurwitz, and W. C. Madow. 1953. Sample Survey Methods and Theory. Vol. 2, Wiley.

Kohler, A. C., D. N. Fitzgerald, R. G. Halliday, J. S. Scott and A. V. Tyler. 1970. Length-weight relationships of marine fishes of the Canadian Atlantic region. Fish. Res. Board Can. Tech. Rep. No. 164.

TABLE 1. Samples from two cod landings.

Length-cm.	Large		Medium		Small		wt-lbs
	1st.	2nd.	1st.	2nd.	1st.	2nd.	from key
40							1.05
43						1	1.31
46			6			2	1.61
49			13	2		18	1.94
52			27	13		41	2.32
55			73	26		21	2.75
58			76	28		2	3.23
61			35	20			3.76
64			17	15			4.34
67			15	9			4.98
70			10	3			5.69
73			3	4			6.46
76			3				7.29
79		2	1				8.20
82		1	1				9.17
85		9					10.22
88	2	8					11.35
91	4	8					12.56
94	2	22					13.85
97	1	11					15.22
100	1	14					16.69
103	1	11					18.25
106	3	3					19.90
109	2	3					21.65
112	1						23.49
115	1	1					25.47
118	1						27.50
121							29.66
124		2					31.93
127							34.32
130	1						36.82
Total	20	95	280	120	0	85	
Sample Wt. (measured) -lbs.	350	1600	875	400	0	200	
Total wt. landed	7430	38520	60340	77920	0	11380	