International Commission for the Northwest Atlantic Fisheries

Serial No. 3343
(D.c.9)

ICNAF Res.Doc. 74/106

ANNUAL MEETING - JUNE 1974

# A probability model for commercial catch sampling

by

W. G. Doubleday

Fisheries and Marine Service,
Biological Station,
St. Andrews, N. B.
Canada.

Abstract:

A general design for commercial groundfish survey sampling is derived and analysed. The relation of the model to current Canadian groundfish sampling practice is discussed.

## Introduction:

This document is a preliminary report concerning work in progress on the evaluation of the Canadian groundfish sampling scheme.

Currently the selection of landings to be sampled and of a sample from a landing is a haphazard process guided largely by convenience. The basic sampling unit is a box of fish. This is selected by a fish plant employee at his convenience and presented to the sampler who may accept or reject it. The landings to be sampled are selected by the sampler who must travel from plant to plant and can only sample when fish are being unloaded from boats.

Earlier authors, Brennan (1) and Gulland (3), have assumed that the basic sampling unit is a single fish and that the number of fish in a given length category found in a box of fish has the binomial distribution. Since it is the sampling procedure that is under investigation, such an assumption is undesirable. Also, it is preferable to consider all length classes together instead of separately as has been done since the observed numbers in different length classes in a box of fish cannot be statistically independent (a box of large fish cannot also be a box of small fish).

Currently, the sampling of commercial landings is aimed at providing estimates of numbers of fish caught in all possible year classes for an ICNAF division and (most frequently) a three month period for each species of interest. These estimates are the raw material for virtual population analyses and catch per unit effort studies.

It was decided to construct a probability model similar in form to the current haphazard scheme. There were two aims in view. One was to determine whether such a scheme could be implemented at reasonable cost, and the other was to form guidelines for the allocation of sampling manpower in the current scheme.

The Model:

The proposed model is a two-way stratified, two-stage cluster sampling procedure. It is possible to simplify the structure by removing one or both of the stratifications for applications to other sampling problems. The population consists of all the boxes of fish landed from one ICNAF division and one three month period of a single species.

The population is divided into major strata consisting of all landings for a particular gear in a time period. The three month period must be subdivided into time periods short enough that the number of landings within a period can be predicted at the beginning of the period so that a random sample of those landings may be chosen. From each stratum, a simple random sample of landings is chosen.

The landings are subdivided into sub-strata consisting of all boxes landed in a market category. From each sub-stratum of a sampled landing, a simple random sample of boxes of fish is chosen and all fish in a box are measured.

Note that all strata are represented in the sample and that all sub-strata of a sampled landing are represented in the sample. Samples from different strata are assumed to be drawn independently. The use of boxes is not essential; the landings could be divided into any collection of equal, non-overlapping volumes and single volume units could be the sampling units.

The conversion of lengths to ages is accomplished with an age length key $\hat{X}$ : $\hat{x}_{st}$ = estimated proportion of fish of length class t that belong to age class s. (The total number of age classes is S and of length classes is T.) $\hat{X}$ is assumed to be distributed independently of the length samples. $E[\hat{X}] = X$, the true age-length relationship of the population. The columns of the matrix $\hat{X}$ are assumed to be mutually independent and the within column dispersion matrix for column t is $\Sigma_t$.

Notation

| | No. in sample | No. in population | subscript |
|---|---|---|---|
| Strata (time period & gear) | G | G | i |
| Clusters (landings | $\ell_i$ | $L_i$ | j |
| Sub-strata (market cat.) | $c_{ij}$ | $C_{ij}$ | k |
| Units (boxes) | $b_{ijk}$ | $B_{ijk}$ | $\ell$ |
| Observed variable | $n_{ijk\ell}$ | $N_{ijk\ell}$ | |

The weight of the contents of a box is $W_{ijk\ell}$ and the total weight of a substratum is $W_{ijk}$.

The observed variable is a column vector consisting of the numbers of fish in the various length (or length-sex in the case of flatfish) categories.

Let $B_i = \sum_{j=1}^{L_i} B_{ij} = \sum_{j=1}^{L_i} \sum_{k=1}^{C_{ij}} B_{ijk}$

$$\underset{\sim}{N} = \sum_{i=1}^{G} \underset{\sim}{N}_i = \sum_{i=1}^{G}\sum_{j=1}^{L_i} \underset{\sim}{N}_{ij} = \sum_{i=1}^{G}\sum_{j=1}^{L_i}\sum_{k=1}^{C_{ij}} \underset{\sim}{N}_{ijk} = \sum_{i=1}^{G}\sum_{j=1}^{L_i}\sum_{k=1}^{C_{ij}}\sum_{\ell=1}^{B_{ijk}} \underset{\sim}{N}_{ijk\ell}$$

The estimation of the true age composition vector $\underset{\sim}{a}$ (numbers of fish landed in the various age categories) of the population is as follows:

$\underset{\sim}{a} = X\underset{\sim}{N}$ is estimated by

$\hat{\underset{\sim}{a}} = \hat{X}\hat{\underset{\sim}{N}}$, where $\hat{\underset{\sim}{N}}$ is calculated in the following stages:

1. Within Sub-strata: $\hat{\underset{\sim}{N}}_{ijk} = \frac{B_{ijk}}{b_{ijk}} \sum_{\ell=1}^{b_{ijk}} \underset{\sim}{n}_{ijk\ell}$

   Note that scalar multiplication is not given a special operation symbol since the meaning is always clear from the context.

2. Within Clusters: $\hat{\underset{\sim}{N}}_{ij} = \sum_{k=1}^{C_{ij}} \hat{\underset{\sim}{N}}_{ijk}$

3. Within Strata: $\hat{\underset{\sim}{N}}_i = \frac{L_i}{\ell_i} \sum_{j=1}^{\ell_i} \hat{\underset{\sim}{N}}_{ij}$

4. Whole Population: $\hat{\underset{\sim}{N}} = \sum_{i=1}^{G} \hat{\underset{\sim}{N}}_i$

Thus $\hat{\underset{\sim}{N}} = \sum_{i=1}^{G} \frac{L_i}{\ell_i} \sum_{j=1}^{\ell_i}\sum_{k=1}^{C_{ij}} \frac{B_{ijk}}{b_{ijk}} \sum_{\ell=1}^{b_{ijk}} \underset{\sim}{N}_{ijk\ell}.$

Expectations:

As a straightforward generalization of standard univariate sampling theory (e.g. Hansen Hurwitz, and Madow (4)) it can be shown that the estimators at each stage are unbiased.

$$E\left[\hat{\underset{\sim}{N}}_{ijk}\right] = \underset{\sim}{N}_{ijk}$$

$$E\left[\hat{\underset{\sim}{N}}_{ij}\right] = \underset{\sim}{N}_{ij}$$

$$E\left[\hat{\underset{\sim}{N}}_{i}\right] = \underset{\sim}{N}_{i}$$

$$E\left[\hat{\underset{\sim}{N}}\right] = \underset{\sim}{N}$$

$$E\left[\hat{\underset{\sim}{a}}\right] = E\left[\hat{X}\right] E\left[\hat{\underset{\sim}{N}}\right] = \underset{\sim}{a}$$

Dispersion Matrices:

The multivariate generalization of a variance is a dispersion (variance-covariance)matrix. The following formulae can be derived by calculating a typical diagonal term (variance) and a typical off-diagonal term (covariance).

1. $$\text{Var}\left[\hat{N}_{ijk}\right] = \frac{B_{ijk}^2}{b_{ijk}} \left(1 - f_{ijk}\right) S_{ijk}^2$$

where $f_{ijk} = b_{ijk} / B_{ijk}$

and

$$S_{ijk}^2 = \sum_{\ell=1}^{B_{ijk}} \frac{\left(\underset{\sim}{N}_{ijk\ell} - \underset{\sim}{N}_{ijk} / B_{ijk}\right)\left(\underset{\sim}{N}_{ijk\ell} - \underset{\sim}{N}_{ijk} / B_{ijk}\right)^T}{B_{ijk} - 1}$$

2. $$\text{Var}\left[\hat{\underset{\sim}{N}}_{ij}\right] = \sum_{k=1}^{C_{ij}} \text{Var}\left[\hat{\underset{\sim}{N}}_{ijk}\right]$$

$$= \sum_{k=1}^{C_{ij}} \frac{B^2_{ijk}}{b_{ijk}} \left(1 - f_{ijk}\right) S^2_{ijk}$$

3. $$\text{Var}\left[N_i\right] = \frac{L_i^2}{\ell_i} \left\{ \left(1 - f_i\right) S_{i1}^2 + L_i^{-1} \sum_{j=1}^{L_i} \sum_{k=1}^{C_{ij}} \frac{B^2_{yk}}{b_{ijk}} \left(1-f_{ijk}\right) S^2_{ijk} \right\}$$

where $f_i = \ell_i / L_i$

and $$S^2_{i1} = \sum_{j=1}^{L_i} \frac{\left(\underset{\sim}{N}_{ij} - \underset{\sim}{N}_i/L_i\right)\left(\underset{\sim}{N}_{ij} - \underset{\sim}{N}_i/L_i\right)^T}{\left(L_i - 1\right)}$$

4. $$\text{Var}\left[\hat{\underset{\sim}{N}}\right] = \sum_{i=1}^{G} \text{Var}\left[N_i\right]$$

$$= \sum_{i=1}^{G} \frac{L_i^2}{\ell_i} \left\{ \left(1-f_i\right) S^2_{i1} + L_i^{-1} \sum_{j=1}^{L_i} \sum_{k=1}^{C_{ij}} \frac{B^2_{ijk}}{b_{ijk}} \left(1-f_{ijk}\right) S^2_{ijk} \right\}$$

$$= \underset{\sim}{\Sigma}^*$$

Note: Dispersion matrices are distinguished from summation signs by a vertical bar ($\underset{\sim}{\Sigma}$ vs $\Sigma$).

5. Since $\hat{\underset{\sim}{a}}$ is the matrix product of two estimators, an exact variance formula is complicated. However, a second order approximation is available.

$$\text{Var}\left[\hat{\underset{\sim}{a}}\right] \underset{\sim}{\approx} \underset{\sim}{\Sigma}^*$$

where entry s, s of $\underset{\sim}{\Sigma}^*$ is $\sum_{t=1}^{T} \underset{\sim}{N}(t)\, \Sigma_t\, (s,s)\, \underset{\sim}{N}(t) + X(s)\, \underset{\sim}{\Sigma}^*\, X^T(s)$

, $X(s)$ represents row s of the matrix $\underset{\sim}{X}$,

and entry s, s' of $\underset{\sim}{\Sigma}^*$ is $\sum_{t=1}^{T} \underset{\sim}{N}(t)\, \underset{\sim}{\Sigma}_t\, (s,s')\, \underset{\sim}{N}(t) + X(s)\, \underset{\sim}{\Sigma}^*\, X^T(s')$

If the $B_{ijk}$ were not known, but all boxes in a substratum had constant weight $W_{ijk}$ then $B_{ijk} = W_{ijk}/w_{ijk}$ defines them.

As was described in Doubleday [2], estimation procedures at the St. Andrews biological station involve a length-weight key $\hat{\underset{\sim}{z}}$, a column vector, whose t'th entry is the estimated average round weight of all fish in length group t in the population (as defined above). We shall assume that $E[\hat{\underset{\sim}{z}}] = \underset{\sim}{z}$ the true length-weight key for the population and that $\text{Var}[\hat{\underset{\sim}{z}}] = \Sigma_z$. If the key is determined by a regression of log (weight) on log (length) as is often the case, $E[\hat{\underset{\sim}{z}}] \neq \underset{\sim}{z}$, but the bias is small and we shall not investigate it further.

The estimation procedure is identical to the earlier case except at the first stage:

$$\hat{N}_{ijk} = W_{ijk} \sum_{\ell=1}^{b_{ijk}} N_{ijk} \Big/ \hat{\underset{\sim}{z}}^T \sum_{\ell=1}^{b_{ijk}} \underset{\sim}{n}_{ijk}$$

The new $\hat{N}_{ijk}$ is a ratio estimator of $N_{ijk}$. The bias of this estimator was examined in a univariate manner in Doubleday [2]. In vector notation, we have the following second order approximation to bias.

$$E[\hat{N}_{ijk}] \approx \left\{ \frac{W_{iik}\, N_{ijk}}{\underset{\sim}{z}^T N_{ijk}} \left( 1 + b_{ijk}^{-1} B_{ijk}^2 (1-f_{ijk}) \frac{\underset{\sim}{z}^T S^2_{ijk}\, \underset{\sim}{z} + N^T_{ijk}\, \Sigma_z\, N_{ijk}}{(\underset{\sim}{z}^T N_{ijk})^2} \right) - \frac{W_{ijk}\, b_{ijk}^{-1} B_{ijk}^2 (1-f_{ijk})\, S^2_{ijk}\, \underset{\sim}{z}}{(\underset{\sim}{z}^T N_{ijk})^2} \right\}$$

It is small comfort that the bias is proportional to $b_{ijk}^{-1}$ since $b_{ijk}$ is usually 1 or 2 in practice. If $f_k$ is negligible and $N_{ijk\ell} = X_{ijk\ell} N_{ijk}$ (scalar multiple of its mean) then $S^2_{ijk} = S_{x2}\ N_{ijk}\ N^T_{ijk}$ and the bias is zero. If $\underset{\sim}{z}^T N_{ijk\ell} = \text{const.}$ then the bias is again zero.

If we write $E[\hat{N}_{ijk}] = N_{ijk} + \Delta_{ijk}$ then

the bias in the estimator of $\underset{\sim}{a}$ can be calculated

$$E[\hat{\underset{\sim}{a}}] = \underset{\sim}{a} + X \sum_{i=1}^{G} \frac{\ell_i}{L_i} \sum_{j=1}^{L_i} \sum_{k=1}^{C_{ij}} \underset{\sim}{\Delta}_{ijk}$$

Thus, the biases are summed in the estimation of $\underset{\sim}{a}$. One would expect most of the $\underset{\sim}{\Delta}_{ijk}$ to point in the same general direction. The approximation to bias is based on the assumption that the coefficient of variation of the denominator is small (0.1 is reasonable).

It is also possible to give second order approximations to the (mean squared error) dispersion matrices, although the expressions are much more complicated than those given above.

$$\text{Var}\,[\hat{N}_{ijk}] \approx \left(\frac{W_{ijk}\,B_{ijk}}{z^T N_{ijk}\,b_{ijk}}\right)^2 \left\{\frac{b^{-1}_{ijk}\,B^2_{ijk}(1-f_{ijk})\,z^T S^2_{ijk}\,z}{(z^T N_{ijk})^2}\right.$$

$$\left. + \frac{N^T_{ijk}\,\Psi_z\,N_{ijk}}{}\right) N_{ijk}\,N^T_{ijk}\,\frac{b^2_{ijk}}{B^2_{ijk}}$$

$$+ b_{ijk}\,(1-f_{ijk})\,S^2_{ijk}$$

$$- \frac{b_{ijk}\,(1-f_{ijk})\,N_{ijk}\,z^T\,S^2_{ijk}}{z^T N_{ijk}}$$

$$\left. - \frac{b_{ijk}\,(1-f_{ijk})\,S^2_{ijk}\,z\,N^T_{ijk}}{z^T N_{ijk}}\right\} = \Psi^1_{ijk}$$

If $k^T n_{ijk\ell}$ is constant (constant weight), this expression reduces to $\left(\frac{W_{ijk}\,B_{ijk}}{z^T N_{ijk}\,b_{ijk}}\right)^2 b_{ijk}(1-f_{ijk})\,S^2_{ijk}$

If $n_{ijk\ell} = x_{ijk\ell}\,N_{ijk}$, then the variance (although not the approximation) is zero.

If $\frac{W_{ijk}}{z^T N_{ijk}} \approx 1$, then approximate mean squared errors may be calculated to be.

1. $\text{Var}\,[\hat{N}_{ijk}] \approx \Psi_{ijk}$

2. $\text{Var}\,[\hat{N}_{ij}] \approx \sum_{k=1}^{C_{ij}} \Psi^1_{ijk}$

3. $\text{Var}\,[\hat{N}_i] \approx \frac{L^2_i}{\ell_i}\left\{(1-f_i)\,S^2_{i1} + L_i^{-1}\sum_{j=1}^{L_i}\sum_{k=1}^{C_{ij}}\Psi^1_{ijk}\right\}$

$$4.\quad \mathrm{Var}\,[\hat{\underset{\sim}{N}}] \approx \sum_{i=1}^{G} \frac{L_i^2}{\ell_i} \left\{ (1-f_i)\, S^2_{i1} + L_i^{-1} \sum_{j=1}^{L_i} \sum_{k=1}^{C_{ij}} \ddagger^1_{ijk} \right\}$$

The approximation to Var (a) is similar in form to the earlier case and is not repeated.

If the size of samples permitted, the various parameters (dispersion matrices) could be estimated using the corresponding sample values.

Practical Observations:

Now that a probability sampling design has been developed, it is possible to compare it with the current scheme and to examine the practical difficulties of implementation.

The first observation is that a large part of some populations is hardly ever sampled. The landings of the mobile fleet in Quebec and the inshore fisheries are rarely sampled. Thus, the selection of landings to sample is far from ideal.

The next observation is that the first and last few boxes of fish in a landing are never sampled. This means that any estimate of a within landings variance is likely to be too small, and that serious biases in numbers at length may exist in some instances.

The third observation is that a market category from a landing is usually represented in the sample by a single box of fish. This means that it is impossible to estimate even roughly the within landings variance.

The fourth observation is that the landing weights $W_{ijk}$ currently used are usually the nominal landing weights calculated at the fish plants as the product of $B_{ijk}$ times a nominal box weight multiplied by a correcting factor to change gutted weight to round weight. The length-weight key requires round weights since it is based on research vessel catches. This process generates unknown biases.

The fifth observation is that even if the current sampling effort were evenly distributed, it would be impossible to obtain more than one sample from a stratum in most populations.

These observations are quite general since detailed figures are not yet fully compiled. However, the available tabulations are sufficient to raise the question of whether it is too ambitious to implement a probability sampling scheme aimed at producing both estimates and reliable confidence intervals without much greater resources. A random sample of size three or four from even a moderately variable population has a sampling error greater than the bias involved in selecting a representative sample from the middle of the range of variation in the population. The current scheme is intended to take at least one sample from each major segment of the populations and to weight these observations according to the size of the corresponding segment. This is a reasonable objective when resources are so limited.

Conclusion:

A probability sampling scheme has been developed to form a framework for study of the current haphazard scheme and to investigate the costs of implementation. The model has demonstrated some inadequacies in the current Canadian commercial groundfish sampling scheme. Further experimentation using the new sampling design on a small scale is necessary before the costs of implementation can be determined. It is evident, however, that several times the current resources will be necessary for implementation.

## REFERENCES

(1) Brennan, J. A. (1974) Preliminary Evaluation of the Present U.S.A. Sampling Scheme of Yellowtail Flounder for Estimating the Number at Age in the Catch Landed. ICNAF Res. Doc. 74/29.

(2) Doubleday, W. G. (1974) Bias in two length frequency formulae. ICNAF Res. Doc. 74/26.

(3) Gulland, J. A. (1955) Estimation of Growth and Mortality in Commercial Fish Populations. Fish. Invest. II, Vo. XVIII, No. 9.

(4) Hansen, M. H., W. N. Hurwitz, and W. G. Madow (1953) Sample Survey Methods and Theory. John Wiley & Sons, New York.