INTERNATIONAL COUNCIL FOR THE
EXPLORATION OF THE SEA

C.M. 1978/D:2
Statistics Committee
Ref.: Pelagic Fish (Northern) Committee.

STRATIFICATION AFTER SELECTION OF THE SAMPLE

by

H.B. Becker
Netherlands Institute for Fishery Investigations,
Haringkade 1, P.O. Box 68, IJmuiden, The Netherlands.

International Council for the
Exploration of the Sea

C.M. 1978/D:2
Statistics Committee
Ref.: Pelagic Fish
(Northern) Committee

STRATIFICATION AFTER SELECTION OF THE SAMPLE

by

H.B. Becker

Netherlands Institute for
Fishery Investigations
Haringkade 1, P.O. Box 68,
IJmuiden, The Netherlands.

Abstract

In constructing strata boundaries the best characteristic is clearly the frequency distribution of Y itself. The next best is presumably the frequency distribution of some quantity which is highly correlated with Y. In geographical stratification the problem is to select variables that have high correlation with the sampled variable, such as temperature, depth, salinity etc. No high correlations with these auxiliary variates were found. Therefore the gains from geographical stratification are likely to be rather modest. In this paper a more mathematical approach has been described. By this method stratification takes place after selection of the random sample. A situation of almost proportional stratification has been achieved, improving the precision of the estimated mean number per haul, provided that the effects of errors in the weights $\frac{N_h}{N}$ can be neglected.

1. Introduction.

In sampling a heterogeneous population stratification may produce a gain in the precision of the estimate of the mean number of fish per haul. For a heterogeneous population like the number of fish per haul taken from a certain area this gain can be rather high. Such a population is therefore divided in subpopulations (strata) which are internally homogeneous; i.e. the variation in the number of fish per haul in each stratum is rather small. The difficulty is that we don't know the stratum to which a haul belongs before the hauls have been made. This paper deals with stratification after the sample has been selected.

2. The method.

Let us consider for example the sample frequency distribution of the number per haul from the North Sea Young Herring Survey. The distribution is very skewed and suitable for stratification.

Given the number of strata, it is possible to construct strata boundaries from this frequency distribution by using the cumulative $\sqrt{f}$-rule of Dalenius and Hodges (Cochran 1963).

In our situation it is most convenient to construct 3 strata.

The hauls in a random sample taken during a further survey are now classified into the 3 strata.

The following notation is used:

for stratum h (h = 1, 2, 3)

N_h	total number of hauls
n_h	number of hauls in the sample
Y_{hi}	number of fish in haul i
$\frac{N_h}{N}$	stratum weight
$\bar{Y}_h = \sum_{i=1}^{n_h} \frac{Y_{hi}}{n_h}$	sample mean

The mean number per haul for the whole area is now estimated by the weighted mean

$$\bar{Y}_w = \sum_{h=1}^{3} \frac{N_h}{N} \bar{Y}_h$$

The weights $\frac{N_h}{N}$ are not known, but can be estimated fairly well

by the sampling proportion $\frac{n'_h}{n'}$ of the sample used for constructing the 3 strata.

According to Desabie (1971) two aspects can be considered:

1. The precision of the weighted estimator $\bar{Y}_w$ before the sample has been taken; thus before the allocation of the number of hauls to the strata took place. The number of hauls n_h are still random variables.

2. The precision of $\bar{Y}_w$ after the sample has been taken; the values of n_h are thus known. This is the normal situation in stratified sampling where the number of hauls n_h (h = 1, 2, 3) have been chosen before.

In the latter case, the variance of $\bar{Y}_w$, given n_1, n_2 and n_3, is estimated by

$$v\,(\bar{Y}_{w/n_1,\ n_2,\ n_3}) = \sum_{h=1}^{3} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h}$$

$$\approx \sum_{h=1}^{3} \left(\frac{n'_h}{n'}\right)^2 \frac{s_h^2}{n_h}$$

where

$$s_h^2 = \frac{1}{n_h - 1} \sum_{h=1}^{n_h} (Y_{hi} - \bar{Y}_h)^2$$

It is assumed that the effect of errors in the weights $\frac{N_h}{N}$ can be ignored.

In order to determine the precision of the estimator $\bar{Y}_w$ before the sample has been taken, the number of hauls n_h now have to be considered as random variables. The estimated variance of $\bar{Y}_w$ is a function of $\frac{1}{n_h}$ and can be written as

$$v\,(\bar{Y}_w) = \sum_{h=1}^{3} \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h} - \sum_{h=1}^{3} \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{N_h}$$

The expected value of $\frac{1}{n_h}$ can be given by

$$E\left(\frac{1}{n_h}\right) = \frac{N}{N_h} \frac{1}{n} + \frac{N}{N_h} \frac{N - N_h}{N_h} \frac{1}{n^2} + \cdots\cdots$$

Hence, neglecting terms of order $\frac{1}{n^2}$, the expected value of the estimated variance of $\bar{Y}_w$ is given by

$$E\ v\ (\bar{Y}_w) = (1 - \frac{n}{N})\ \frac{1}{n} \sum_{h=1}^{3} \frac{N_h}{N}\ s_h^2 + \frac{1}{n^2} \sum_{h=1}^{3} (1 - \frac{N_h}{N})\ s_h^2$$

The factor $1 - \frac{n}{N}$, which is the finite population correction, can be omitted because the sampling fraction is very small, while for S_h^2 its unbiased estimator s_h^2 will be substituted.

The first term in this expression is the value of the variance of $\bar{Y}_{st}$ for proportional stratification. This can easily be understood since a simple random sample distributes itself approximately proportionally among strata. The second term represents the increase in variance arising from the distribution of n_h over the strata which is not proportionally. This increase in variance will be small if n_h is reasonably large (e.g. $n_h > 20$).

In other words: stratification a posteriori is almost as precise as proportional stratified sampling, provided that the number of sampling units allocated to each stratum is reasonably large and the effects of errors in the weights $\frac{N_h}{N}$ can be ignored.

The 90% confidence limits for $\bar{Y}_w$ are calculated by the formula

$$\bar{Y}_w \pm t \sqrt{var\ (\bar{Y}_w)}$$

where t is the 90% value of Student's t-distribution with n-3 degrees of freedom and

$$var(\bar{Y}_w) = \frac{1}{n} \sum_{h=1}^{3} \frac{n'_h}{n'} s_h^2$$

3. Application to the North Sea Young Herring Survey in 1977.

From the data of the young herring survey in 1976 3 strata have been constructed by using the cumulative $\sqrt{f}$ - rule.

All the hauls taken from the Ices rectangles in the North Sea area between 52°00'N and 57°30'N have been considered.

It is found:

Stratum I:	0-100	n'_1 = 149 hauls
Stratum II:	100-500	n'_2 = 34 hauls
Stratum III:	> 500	n'_3 = 31 hauls

Next go back to the survey in 1977.

The individual hauls are classified into the three strata as given in table 1.

The mean number per haul and its confidence limits are calculated as follows:

$$\bar{Y}_w = \frac{149}{214} \times 17.70 + \frac{34}{214} \times 264.46 + \frac{31}{214} \times 1657.82 = 294.5$$

$$\text{Var}\,(\bar{Y}_w) = \frac{149}{214} \times \frac{724.5}{218} + \frac{34}{214} \times \frac{12801.4}{218} + \frac{31}{214} \times \frac{2885325.9}{218} = 1929$$

90% confidence limits of $\bar{Y}_w$: $294.5 \pm 1.65 \times 43.92 = 294.5 \pm 72.5$

For comparison also the 90% confidence limits have been calculated for the overall mean $\bar{Y}$ of the whole sample and the stratified mean $\bar{Y}_{st}$ by the method of geographical stratification.
It is found:

$$\bar{Y} \pm 1.65 \sqrt{\text{var}\,(\bar{Y})} = 350 \pm 104$$

and

$$\bar{Y}_{st} \pm 1.65 \sqrt{\text{var}\,(\bar{Y}_{st})} = 303 \pm 115$$

4. Acknowledgments

The author is grateful to Mr. Aksland, University of Bergen, for his valuable comments.

References

Cochran, W.G., (1963) - Sampling Techniques
John Wiley and Sons, New-York, London.

Desabie, J., (1971) - Théorie et practique des sondages;
Dunod, Paris.

Table 1 - Classification of the individual hauls from the 1977 survey into the 3 strata.

Individual hauls per stratum									
Stratum I					Stratum II		Stratum III	
24	0	2	7	66	212	304	2261	2048
0	0	4	80	4	472	206	976	1497
2	2	21	19	2	338	315	992	760
0	42	0	85	31	106	224	661	792
6	0	0	20	70	245	116	1065	1080
0	0	8	82	62	498	168	762	608
0	16	0	0	62	256	157	3902	580
56	0	0	0	58	224	240	3470	800
0	1	8	86	22	180	244	1725	2640
10	1	3	70	0	127	472	752	1612
8	0	0	3	71	396	334	876	
0	23	0	54	27	264	118	744	
2	4	4	26	83	234	214	677	
0	44	0	65	75	172		1304	
0	0	3	1	26	468		2982	
0	0	0	14	2	306		2949	
0	2	0	30	4	200		1220	
26	0	4	2	26	105		4816	
53	4	0	0	4	111		1208	
71	0	3	0	22	151		1420	
2	20	0	0	14	276		9792	
0	2	1	4	16	273		509	
1	1	0	0	10	266		636	
0	2	19	82	60	438		1650	
1	1	4	95	64	262		513	
2	2	0	33	98	271		1230	
2	0	1	24	74	432		952	
0	2	1	4		448		536	
$n_1 = 139$					$n_2 = 41$		$n_3 = 38$	
$\bar{Y}_1 = 17.70$					$\bar{Y}_2 = 264.46$		$\bar{Y}_3 = 1657.82$	
$s_1^2 = 724.5$					$s_2^2 = 12801.4$		$s_3^2 = 2885325.9$	