Sampling Requirements for the Adult Fish Survey

SUSAN PICQUELLE[1]
Southwest Fisheries Center,
National Marine Fisheries Service, NOAA,
P.O. Box 271, La Jolla, CA 92038

ABSTRACT

The question of how many fish should be collected to estimate population fecundity parameters is addressed. The goal is to determine the optimal combination of the number of stations to occupy (n) and the number of fish to subsample per station (m) that will provide the minimum variance of parameter estimates. The collection of mature fish for this purpose is considered to be a two-stage sample design in which the total number of fish processed is equal to n times m. In most cases, the subsample of fish taken from a trawl station can be treated as a cluster sample, and a simple relationship between variance of the parameters and the sample sizes, n and m, may be established. Spawning fraction, the parameter with the largest relative variance, is chosen to evaluate alternative combinations of n and m. In terms of the total number of fish processed, it is generally more efficient in a statistical sense to occupy more trawl stations and to subsample fewer fish per station than vice-a-versa. If the major cost of collecting data is associated with ship operations, then it is cheaper to occupy fewer trawl stations and subsample more fish per station. The equations and Figure 1, based on northern anchovy data from California, provide criteria for determining the combination of n and m which achieves desired levels of variance.

INTRODUCTION

The number of independent observations directly impacts the estimated variance of a parameter; the higher the number of observations, the smaller the variance. A trawl survey is usually a two-stage sample design; hence the number of observations is determined by the number of trawls taken (n) and the number of fish subsampled from each trawl (m). However, for a fixed total sample size (nm), varying combinations of n and m will produce varying values of the estimated variance because the fish within a trawl are typically more similar than fish between trawls, that is, fish within trawls are positively correlated and hence are not independent observations. For example, the intratrawl correlation coefficient for female weight data collected during the 1980 survey was 0.60. Thus it is advantageous to find the optimal combination of n and m that will produce the minimum variance.

SAMPLE SIZE RELATIONSHIPS

In two-stage sampling, the estimates of the population mean and variance are (Cochran 1963),

$$\bar{\bar{x}} = \sum_{i=1}^{n} \frac{\bar{x}_i}{n} \quad \text{where } \bar{x}_i = \sum_{j=1}^{m} \frac{x_{ij}}{m} \tag{1}$$

$$\text{and } \hat{var}(\bar{\bar{x}}) = (1 - f_1)\frac{s_1^2}{n} + f_1(1 - f_2)\frac{s_2^2}{nm} \tag{2}$$

where $s_1^2 = \dfrac{\sum_{i=1}^{n} (\bar{x}_i - \bar{\bar{x}})^2}{n - 1}$ = intertrawl component of variance,

$s_2^2 = \dfrac{\sum_{i=1}^{n}\sum_{j=1}^{m} (x_{ij} - \bar{x}_i)^2}{n\,(m - 1)}$ = intratrawl component of variance,

$f_1 = n/N$, where N is the total number of stations, and
$f_2 = m/M$, where M is the total number of elements at each station.

(Note: for simplicity's sake, two-stage sampling with stations of equal size is used here for illustration purposes.) Equation (2) relates variance to the values of n and m, and this equation can be used to find the optimal values of n and m which will produce the minimum variance.

But first, some simplifications can be made. In the majority of fisheries surveys, the sampling fraction f_1 is negligibly small so that the value $f_1 = 0$ may be substituted into equation (2). This reduces to

$$\hat{var}(\bar{\bar{x}}) = \frac{s_1^2}{n} \tag{3}$$

and the intratrawl component of variance disappears. Because of this simplification, the trawl sample may be treated like a cluster sample. In cluster sampling, the trawl average, $\bar{x}_1$, is measured without error, i.e., every fish is measured. In two-stage sampling, the trawl average is estimated with error, but in this case that error is negligible and does not impact the total variance estimate.

By considering the subsample of fish taken from a trawl as a cluster sample, a simpler relationship between variance and values of n and m may be established. The intracluster correlation (ρ) may be expressed as a function of the ratio of the cluster-sample variance to the random-sample variance:

[1]Present address: Northwest and Alaska Fisheries Center, National Marine Fisheries Service, NOAA, 7600 Sand Point Way NE, Seattle, WA 98115.

$$\frac{\sigma^2_{\bar{x}_{cluster}}}{\sigma^2_{\bar{x}_{random}}} = 1 + \rho\,(m - 1) \tag{4}$$

where
$$\sigma^2_{\bar{x}_{cluster}} = \frac{\sum_{i=1}^{n} (\bar{x}_i - \bar{\bar{x}})^2}{n\,(n-1)} \tag{5}$$

and
$$\sigma^2_{\bar{x}_{random}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} (x_{ij} - \bar{\bar{x}})^2}{nm\,(nm-1)} \cdot \tag{6}$$

The fish sampled during the trawl survey are used to estimate several parameters, i.e., sex ratio, spawning fraction, fecundity, and female weight. The optimal combination of m and n will not be the same for all parameters. Hence, the parameter with the largest relative variance is chosen to pick the values of m and n. This parameter is spawning fraction for the northern anchovy example.

Spawning fraction is distributed approximately as a binomial distribution. We can simplify equation (4) by substituting the variance of a binomial distribution for the random-sample variance

$$\sigma^2_{\bar{x}_{random}} = \sigma^2_{\bar{x}_{binomal}} = \frac{\bar{\bar{x}}\,(1 - \bar{\bar{x}})}{nm} \cdot \tag{7}$$

Solving equation (4) for m after substituting equation (7) for $\sigma^2_{\bar{\bar{x}}_{random}}$ gives

$$m = \frac{\bar{\bar{x}}\,(1 - \bar{\bar{x}})(1 - \rho)}{n\sigma^2_{\bar{x}_{cluster}} - \bar{\bar{x}}\,(1 - \bar{\bar{x}})\rho} \cdot \tag{8}$$

Values for $\bar{\bar{x}}$, ρ and $\sigma^2_{\bar{\bar{x}}_{cluster}}$ are needed to specify the relationship between m and n. It is more convenient to work with the coefficient of variation (cv) rather than variance directly because one can then specify the desired precision in terms of percent of the parameter rather than in absolute terms. This changes equation (8) to

$$m = \frac{(1 - \bar{\bar{x}})(1 - \rho)}{n\bar{\bar{x}}\,cv^2 - (1 - \bar{\bar{x}})\rho} \tag{9}$$

where
$$cv = \frac{\sigma_{\bar{\bar{x}}_{cluster}}}{\bar{\bar{x}}} \cdot$$

Considering a range of values for the coefficient of variation is useful to see how n and m change as the precision of the estimate of $\bar{\bar{x}}$ changes. It is also worthwhile to consider a range of values for $\bar{\bar{x}}$, because estimates of this are not available until after the survey, but it is possible to specify a range of values within which $\bar{\bar{x}}$ is likely to fall. A value for ρ may be selected using an estimate from previous surveys.

NUMERICAL EXAMPLE

As an example, the attached series of graphs (Fig. 1) was constructed using the estimate of ρ calculated from the 1982 survey data using Equation (4) (ρ = 0.0448). Values of $\bar{\bar{x}}$ range from 0.06 to 0.12 in increments of 0.02, with one graph corresponding to each value of $\bar{\bar{x}}$. On each graph are five lines corresponding to five values of the desired coefficient of variation (0.100 to 0.200). The vertical axis is m, the subsample size; and the horizontal axis is n, the number of trawls.

These graphs do not serve to pinpoint the precise combination of m and n that is optimal. Instead, they illustrate how n and m together determine the precision of the estimate, given the value of the estimate. They suggest general sampling strategies rather than specific criteria.

It is generally more efficient (in terms of number of fish processed) to take more trawls and fewer fish per trawl than vice-a-versa. For example, consider a spawning fraction of 0.10 and a desired coefficient of variation of 0.125. If 80 trawls are taken, then 10 fish should be subsampled, for a total of 800 fish. On the contrary, if only 50 trawls are taken, then 26 fish need to be subsampled, for a total of 1,250 fish. This is true because the fish within each trawl tend to be positively correlated and hence are not independent observations, while the fish between trawls are uncorrelated, thus contributing more information per fish.

COST CONSIDERATIONS

The major cost of data collection is associated with ship operations, not with processing the fish in the laboratory. For a fixed number of fish sampled, it is much cheaper to take fewer trawls and larger subsamples, although this reduces the precision of the estimates. Thus, what is more efficient statistically (many trawls and small subsamples) is the exact opposite of what is more efficient financially (few trawls and large subsamples). If the costs of taking a trawl (C_1) and of processing a fish (C_2) are known, these costs may be incorporated in the relationship between n and m and a new function may be derived which minimizes total cost (C),

$$C = nC_1 + nmC_2 \cdot \tag{10}$$

We wish to minimize C under the constraint of Equation (9).

$$C = nC_1 + nC_2\left[\frac{(1 - \bar{\bar{x}})(1 - \rho)}{n\bar{\bar{x}}\,cv^2 - (1 - \bar{\bar{x}})\rho}\right] \cdot \tag{11}$$

C is minimized by setting $\dfrac{\partial C}{\partial n} = 0$.

This results in the following values for n and m:

$$n = \frac{C_1(1 - \bar{\bar{x}})\rho + (1 - \bar{\bar{x}})\sqrt{C_1C_2\rho(1 - \rho)}}{C_1\bar{\bar{x}}\,cv^2} \tag{12}$$

$$m = \left[\frac{(1 - \rho)C_1}{\rho C_2}\right]^{1/2} \cdot \tag{13}$$

Consider the example presented earlier where $\bar{\bar{x}}$ = 0.10 and the desired coefficient of variation is 0.125 (again using ρ = 0.0448 from the 1982 survey data). Suppose the cost of taking one trawl (C_1) is \$1,000 and the cost of processing one fish (C_2) is \$25. These conditions result in the values minimizing cost of n = 45 and m = 29.

Frequently the number of trawls taken is determined by factors not related to desired precision and not directly controllable, such as the number of days at sea, weather condition, and the success rate of catching the target species. Before a survey begins, only the number of days at sea is known; however, predictions about the other two factors mentioned are available, so that an approximate number of trawls can be deduced. Then the graphs may be used to find the necessary value of m to attain the desired coefficient of variation.

Another consideration of selecting n and m is the expected number of spawning females subsampled in a trawl. If the spawning frac-

tion is 0.10, and only 10 females are subsampled from each trawl, then the expected number of spawning females sampled from each trawl is only 1, based on a binomial distribution where $n = 10$ and $p = 0.10$. This will lead to a high proportion (35% on average) of trawl subsamples with no spawning females, thus inflating the variance. By raising m to 15 females, the average percent of trawl subsamples with no spawning females drops to 21%; for 20 females the percent is 12%.

LITERATURE CITED

COCHRAN, W. G.
1963. Sampling techniques. John Wiley and Sons, N.Y., 413 p.

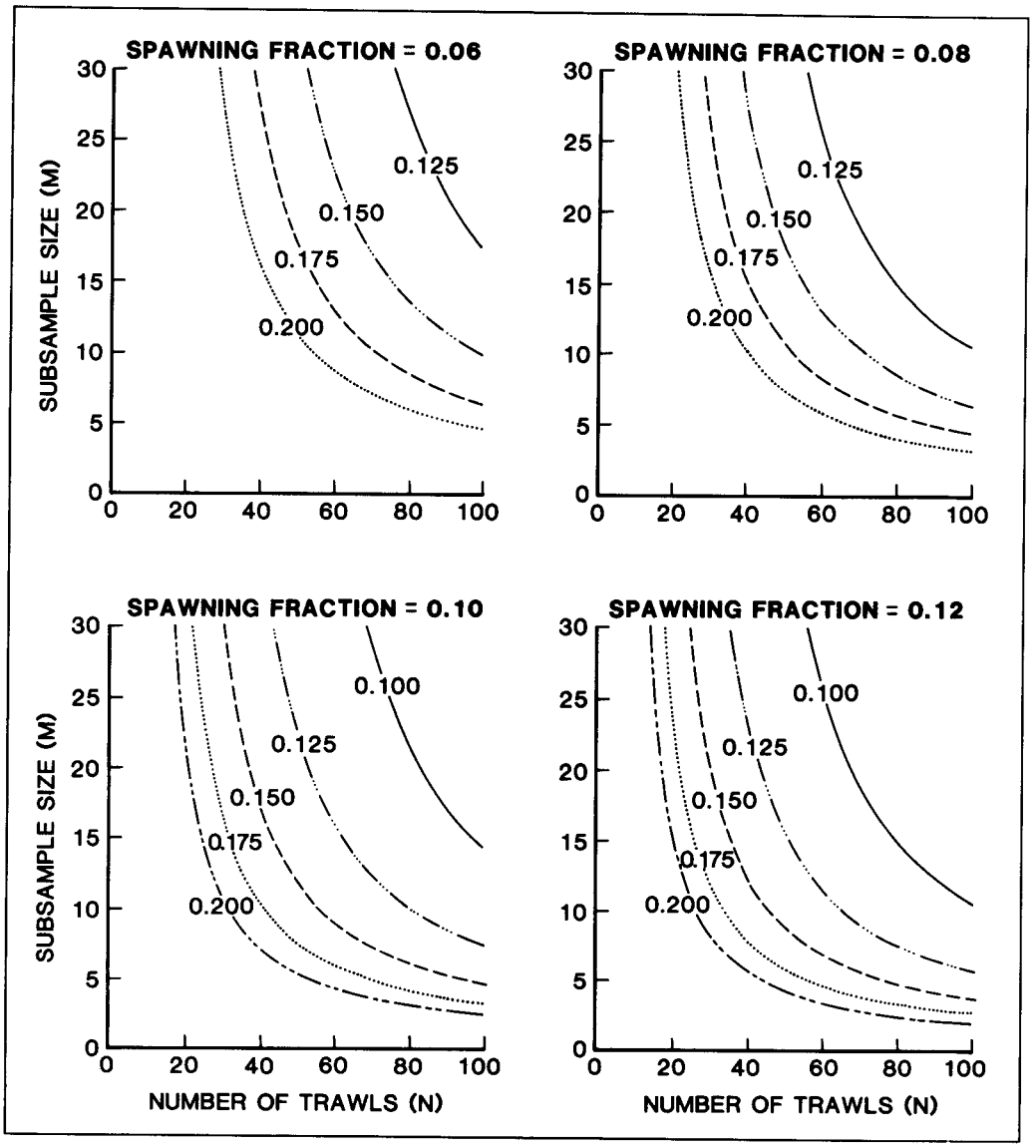

Figure 1.—Relationship between subsample size and number of trawls for a range of values for spawning fraction and coefficient of variation.

UNITED STATES DEPARTMENT OF COMMERCE
National Oceanic and Atmospheric Administration
NATIONAL MARINE FISHERIES SERVICE

Southwest Fisheries Center
P.O. Box 271
La Jolla, California 92038

April 16, 1987

MEMORANDUM FOR: USERS OF THE EGG PRODUCTION METHOD FOR ESTIMATING SPAWNING BIOMASS OF PELAGIC FISH.

FROM: REUBEN LASKER

SUBJECT: ERRATA FOR NOAA TECHNICAL REPORT NMFS 36; "AN EGG PRODUCTION METHOD FOR ESTIMATING SPAWNING BIOMASS OF PELAGIC FISH: APPLICATION TO THE NORTHERN ANCHOVY".

A number of printing errors have been discovered by Dr. Sachiko Tsuji in the published account of the egg production method. These are important and warrant this memo. Please make these corrections in your copy.

p. 5, Abstract, 4th line should read:
"be estimable and spawning rate constant over the field sampling interval."

p. 12, in equation 8, $\hat{\beta}$ should be β.

p. 17, Table 1. on the January line +3.5 should be -3.5.

p. 20, two lines under the formula in the second column, "sample size" should be "sample scale" and σ_i should read σ_{ij}^2. Five lines under the formula "larger observations" should be "bigger scales."

p. 22, 1st para., No. 3 last line should be simulation, not stimulation.

p. 23. 1st para., line 7. "Table 9" should read "Table 6."

p. 44. Temperature table in second column on the page.

The temperatures read	13.9
	13.5
	16.2
The correct temperatures are	13.9
	15.2
	16.2.

p.45. Second column, $y_{i,t,k}$ should read $y_{i,t}$.

p.46 1st Para., line 7, change the word "spawning" to "tows, $\hat{T}$".

p.49. Table 5d. Strike out the words "within or" in the second line of the heading.

p.55. 9th line from the bottom, x_1 should be x_i.

p.56. First. para. second column, sixth line, 26 should read 25.

p.63. Under "Preservation" $Na_2H_2PO_4$ should be Na_2HPO_4.

p.93. In table 1, atretic state e, change > to <.

p.97. In the! formula after the second para. change < to >.

p.98. In the formula in the first column change -Zt to $-Zt_h$.