Limnol. Oceanogr., 31(1), 1986, 211–215

Optimal allocation of effort in studies using the size-frequency method of estimating secondary production

Abstract—Secondary production estimates based on the size-frequency method can be improved through the use of optimal sample allocation without increasing the total sampling effort. We advocate a two-step procedure. First, sampling dates should be chosen to minimize bias. Then, sample sizes on each date should be allocated in such a way as to minimize variance. Calculating the optimal sample allocation requires some prior variance information, but even rather imperfect information will usually result in improved estimates. An example is presented using published data on the mayfly *Ephemerella dorothea*.

An estimate of secondary production is of little value without some idea of the potential bias and variability of the estimate. Estimation of aquatic macroinvertebrate secondary production has received substantial attention in recent years, and several aspects of bias and variability in estimation have been examined. Resh (1979) and Waters (1979) identified biological and mechanical factors which contribute to bias and variability. Krueger and Martin (1980) and Newman and Martin (1983) concentrated on quantifying the sampling variability of production estimates.

Efforts to reduce the error have focused mainly on mechanical or biological problems, such as loss of small animals due to coarse screens, diurnal behavior patterns altering the animals' likelihood of being caught, etc. (Resh 1979; Waters 1979). However, even if these problems were eliminated, secondary production estimates would still often have large sampling variances because they are based on small random samples from the parent population. We describe here a method of allocating effort which can substantially reduce the variance of size-frequency production estimates without increasing the overall study effort.

The size-frequency approach is the method most commonly used to estimate macroinvertebrate production. A concise history and overview of the procedure is given by Krueger and Martin (1980). Krueger and Martin presented equations for estimating the variance of size-frequency production estimates; we have recast these equations so that the optimal allocation problem can be solved in a manner analogous to that used for stratified random sampling (e.g. Cochran 1977). We argue that sampling date selection is important in controlling bias and should be done as the first step in designing a study. Then the number of samples processed from each date should be chosen by optimal sample allocation so as to reduce variance.

Our method requires some prior information about variances on each of the sampling dates, but will still be useful even if the variance information is quite rough (Cochran 1977, p. 115–117; Kish 1965, p. 94–95). We envision two general applications of the technique. Data from previous studies can be used to determine sample sizes to be taken in the field for a future study, or, in the absence of prior information, excess samples can be taken in the field, and a small preliminary sample counted to determine further effort allocation in the lab. We give an example that illustrates the first situation. The second approach is feasible because large numbers of samples can often be collected in the field with little effort; most of the effort in aquatic macroinvertebrate studies is in sorting, identifying, measuring, and enumerating samples in the lab (Resh and Price 1984). The technique is very flexible, and a sound design could incorporate both approaches.

M. Butler stirred our interest in sampling problems in secondary production estimation. R. Newman, P. Wingate and R. Lake provided advice. The comments of anonymous reviewers improved the clarity of the paper.

In the following considerations, each sampling date is indexed by i ($i = 1, 2, \ldots, n$). The (random) samples within each sam-

pling date are indexed by k ($k = 1, 2, \ldots, b_i$). The total number of samples is given by $N = \Sigma b_i$. Each animal belongs to a size class j ($j = 1, 2, \ldots, a$). Y_{ijk} is the number of animals observed on date i in size class j in sample k. The variance of Y_{ijk} will be denoted by σ_{ij}^2. The mean number of animals observed on date i in size class j is

$$\bar{Y}_{ij.} = \sum_{k=1}^{b_i} Y_{ijk}/b_i. \tag{1}$$

The variance of $\bar{Y}_{ij.}$ is

$$v(\bar{Y}_{ij.}) = \sigma_{ij}^2/b_i. \tag{2}$$

An estimate of σ_{ij}^2 is given by the usual formula

$$\hat{\sigma}_{ij}^2 = \frac{\sum_{k=1}^{b_i} (Y_{ijk} - \bar{Y}_{ij.})^2}{b_i - 1} \tag{3}$$

for $b_i > 1$. The mean number of individuals in a size class j during the study is estimated by

$$\bar{Y}_{.j.} = (2D_n)^{-1} \sum_{i=1}^{n-1} (D_{i+1} - D_i) \cdot (\bar{Y}_{ij.} + \bar{Y}_{i+1j.}) \tag{4}$$

where D_i is the number of days between the first sampling date and the ith sampling date ($D_1 = 0$). The variance of $\bar{Y}_{.j.}$ is

$$v(\bar{Y}_{.j.}) = (2D_n)^{-2} \sum_{i=1}^{n} T_i v(\bar{Y}_{ij.}) \tag{5}$$

where

$$T_1 = D_2^2,$$
$$T_2 \text{ to } T_{n-1} = (D_{i+1} - D_{i-1})^2,$$

and

$$T_n = (D_n - D_{n-1})^2.$$

Equation 5 is a reparameterization of Krueger and Martin's (1980) equation 4. (The term $\bar{Y}_{i+1j.}$ in their equation is a typographical error; it should be $\bar{Y}_{1j.}$. Also their D_1 is not needed since it always equals zero by definition.) The production estimate is

$$P_w = a\left[\sum_{j=1}^{a-1} (\bar{Y}_{.j.} - \bar{Y}_{.j+1.})(W_j W_{j+1})^{0.5}\right] + a(\bar{Y}_{.a.} W_a) \tag{6}$$

where W_j is the mean individual weight of an animal in the jth size class. Benke's (1979) correction multiplier of 365/(cohort production interval) is ignored since it cancels out of the optimal allocation calculations. On the assumptions that the size-class mean weights per individual W_j have negligible sampling and measurement errors and that the mean size-class abundances $\bar{Y}_{.j.}$ are uncorrelated, the variance of the production estimate P_w is

$$v(P_w) = a^2 \sum_{j=1}^{a} R_j v(\bar{Y}_{.j.}) \tag{7}$$

where

$$R_1 = W_1 W_2,$$
$$R_2 \text{ to } R_{a-1} = [(W_j W_{j+1})^{0.5} - (W_{j-1} W_j)^{0.5}]^2,$$

and

$$R_a = [W_a - (W_{a-1} W_a)^{0.5}]^2.$$

(Krueger and Martin 1980 discussed the implications of assuming the $\bar{Y}_{.j.}$ values are uncorrelated.) By substituting Eq. 2 into 5 and Eq. 5 into 7 and changing the order of summation, we see that

$$v(P_w) = \sum_{i=1}^{n} V_i/b_i \tag{8}$$

where

$$V_i = a^2 T_i \sum_{j=1}^{a} R_j \sigma_{ij}^2. \tag{9}$$

The term V_i/b_i is the component of variability attributable to the ith sampling date. The primary difference between our variance formula (Eq. 8) and Krueger and Martin's (equation 6) is that ours has been arranged to isolate the daily variance components. Another somewhat technical difference is that Krueger and Martin's equation is given as an estimator; we give ours as a true population variance.

The first step in designing a study to calculate P_w is to select sampling dates that minimize the bias of P_w. The minimization of $v(P_w)$ becomes an issue only after the dates have been selected. We now develop this point. If the true number of animals in size class j at time t is $S_j(t)$, the true mean number of animals in size class j over the entire sampling period is given by

$$D_n^{-1} \int_0^{D_n} S_j(t)\, dt \tag{10}$$

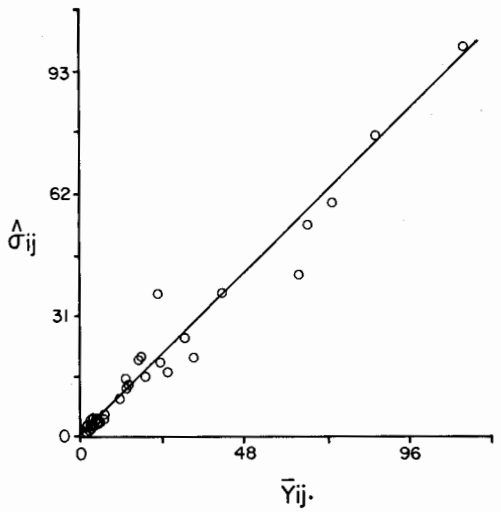

Fig. 1. Estimated standard deviation $\hat{\sigma}_{ij}$ as a function of $\bar{Y}_{ij\cdot}$. Data are from Krueger and Martin (1980, table 1). All zero values of $\bar{Y}_{ij\cdot}$ were dropped from the data set. The line is the least-squares regression model $\hat{\sigma}_{ij} = 0.863\bar{Y}_{ij\cdot}$. ($r^2 = 0.98$, 36 df).

Table 1. Optimal allocation of effort among fixed dates suggested by Krueger and Martin's (1980) data. Total effort is fixed at 130 jars.

	Optimal allocation of samples (b_i) Estimator of σ_y^2		Estimate of daily variance component V_i using $\hat{\sigma}_{ij}^2$
	$\hat{\sigma}_y^2$	$\bar{Y}_{ij\cdot}^2$	
1 Mar	12	12	0.103
23 Mar	14	18	0.134
24 Apr	59	62	2.554
26 May	5	4	0.016
24 Jun	0	0	0
22 Jul	0	0	0
2 Aug	0	0	0
21 Sep	1	0	0
20 Oct	5	3	0.021
29 Nov	4	3	0.012
24 Dec	7	7	0.039
22 Jan	14	11	0.148
28 Feb	9	10	0.064

[assuming that the population is large enough that $S_j(t)$ can be thought of as a continuous function]. Equation 4 can be recognized as an approximation to Eq. 10. Equation 4 assumes that each $S_j(t)$ is linear between the sampling dates; i.e. Eq. 4 performs linear interpolation. Therefore, the bias produced by Eq. 4 is small if the dates are chosen so that the segments of the $S_j(t)$ functions between dates are nearly linear. This means that the sampling dates should be closer together when the slopes of $S_j(t)$ change rapidly. The size classes that are likely to contribute substantially to production should be most influential in determining sampling dates. If the shapes of the $S_j(t)$ functions are not well known, sampling intervals of uniform width should be used.

Krueger and Martin suggested that uniformly spaced intervals minimize $v(\bar{Y}_{j\cdot})$ but generally this is not true. Although not explicitly represented as such, the variances of the estimated daily abundances $v(\bar{Y}_{ij\cdot})$, on which the variances of the mean size class abundances $v(\bar{Y}_{j\cdot})$ are based, are functions of time. If the forms of the $v(\bar{Y}_{ij\cdot})$ functions over time were known, the variance $v(\bar{Y}_{j\cdot})$ would be reduced by concentrating sampling on dates when the $v(\bar{Y}_{ij\cdot})$ functions are small. However, this would not be desirable in general because the resulting values of $\bar{Y}_{j\cdot}$ could be severely biased. As an extreme example, $v(P_w)$ would generally be minimized by sampling only times when animals were absent. The resulting production estimate (i.e. zero) would have a huge bias, but it would have zero variance. Bias reduction alone should determine sampling dates.

The notion of choosing the sampling dates to reduce bias is intuitive. Resh (1979) also suggested that sampling should be most frequent when population sizes and composition are changing most rapidly. Once the sampling dates have been chosen, attention can be directed to selecting the sample sizes on each date (b_i) which minimize the variance of the production estimate $v(P_w)$. With methods similar to those used for stratified random sampling (Cochran 1977), we can show that this occurs when

$$b_i = \frac{NV_i^{0.5}}{\sum_{h=1}^{n} V_h^{0.5}}. \qquad (11)$$

The terms V_i are analogous to stratum variances in stratified random sampling. Optimal allocation will produce large gains in precision when the V_i terms differ considerably (Cochran 1977).

In addition to the sampling dates, Eq. 11

requires estimates of mean weight for each size class (W_j) and the within-sampling-day variance σ_{ij}^2. The former should not be difficult to obtain. The variance σ_{ij}^2 could be estimated as $\hat{\sigma}_{ij}^2$ with Eq. 3 if the necessary data were available. However, there is an alternative if such direct estimates are not available. There is usually a strong relationship between abundance and the within-sampling-day variability. Thus, knowledge of $\bar{Y}_{ij\cdot}$ (estimated daily abundance) can be used to estimate σ_{ij}^2 (Perry 1981).

Application of the method will be demonstrated with data from a mayfly (*Ephemerella dorothea*) population in a southern Minnesota stream sampled by Krueger and Martin (1980). There is a strong linear relationship between the estimated daily abundance $\bar{Y}_{ij\cdot}$ and within-day standard deviation $\hat{\sigma}_{ij}$ (Fig. 1). Such relationships can be exploited to arrive at reasonable guesses of σ_{ij}^2 to substitute into Eq. 11. Only the proportional relationship among the σ_{ij}^2 values is important; the same optimal b_i values would be returned if $c\sigma_{ij}^2$ were substituted into Eq. 11 instead of σ_{ij}^2, since the constants c would all cancel. For example, since the model $\hat{\sigma}_{ij} = \beta \bar{Y}_{ij\cdot}$ appears to fit Krueger and Martin's data well, $\bar{Y}_{ij\cdot}$ could be substituted into Eq. 11 as an estimate of σ_{ij} in this example.

The optimal allocation calculated from Krueger and Martin's table 1 is shown in our Table 1. As in the original study, the total number of samples to be processed is fixed at $N = 130$. Note the similarity in allocation, whether $\hat{\sigma}_{ij}^2$ or $\bar{Y}_{ij\cdot}^2$ is used as an estimate of σ_{ij}^2. Optimal allocation concentrates sampling on dates when biomasses are high, these dates being the ones that contribute most to the variance of P_w.

The sampling scheme suggested by optimal allocation need not be followed rigidly. It may be desirable to have some required fixed minimum sample size on each date. Suppose in the above example it is decided that at least five samples should be taken on each date to ensure that important events in the life cycle are not missed. A compromise between uniform effort sampling and optimal allocation could be made. For example, a sampling regime of 8, 11, 53, 5, 5, 5, 5, 5, 5, 5, 11, 7 (sample sizes ordered by date) would be reasonable. Such compromise regimes usually perform quite well (Cochran 1977; Kish 1965).

The estimated production variance $v(P_w)$ under Krueger and Martin's (1980) uniform effort sampling regime is 0.54 $(g\ m^{-2})^2$. The estimated variance under the compromise regime is 0.20, representing a 63% reduction. This calculated reduction may be overly optimistic because of the errors involved in estimating the σ_{ij}^2, but it gives a general indication of the improvement that can be expected (Cochran 1977, p. 103). Cochran gave the guideline that optimal allocation may not be worthwhile unless a variance reduction >10–20% is expected. Optimal allocation may provide substantial increases in precision even if the sample allocation is considerably different from the unknown true optimum (Cochran 1977; Kish 1965). This means that the estimators of σ_{ij}^2 do not need to be very accurate for optimal allocation procedures to be beneficial.

If one wanted to repeat Krueger and Martin's field sampling, one should use a sampling scheme similar to the compromise regime. However, we had much high quality prior information to design this regime. Suppose one had no prior information to make such design decisions. Must one take massive numbers of samples each time in the field to ensure enough to accommodate the "worst possible daily allocation" likely to be indicated by optimal allocation? For example, Krueger and Martin's "worst possible daily allocation" is 53 samples on 24 April. Practical compromises have to be made. More samples should be taken when biomasses are high. The researcher in the field will often be able to judge after a few samples whether biomasses are high; if they are, more samples should be taken.

It is possible that after a preliminary processing of samples, more effort has been devoted to some dates than is called for by optimal allocation. As a practical solution the date which was most oversampled should be dropped from the allocation calculations and the optimal allocation for the remaining dates should be recomputed. This procedure should be repeated until no oversampled dates remain. A similar procedure

should be used if optimal allocation calls for more samples on a date than are available.

The intended use of a secondary production estimate should dictate how much uncertainty in the estimate is tolerable. In some situations a highly variable, undependable estimate may be worse than no estimate at all. Even with optimal allocation the variance of the production estimate may be unacceptably large. In such a case the total sample size must be increased to reduce the variance.

To calculate the total sample size needed to achieve a desired variance, we must specify beforehand the proportion of the total sample to be taken on each date, denoted as p_i. To find the p_i values, we select any arbitrary total sample size N and calculate the optimal daily sample sizes b_i. Then,

$$p_i = b_i/N. \quad (12)$$

Suppose we decide that the variance of the production estimate should equal some value F. Substituting Eq. 12 into 8 and then rearranging gives

$$N^* = F^{-1}a^2 \sum_{i=1}^{n} \sum_{j=1}^{a} T_i R_j \sigma_{ij}^2/p_i. \quad (13)$$

That is, a total sample size of N^* should yield $v(P_w) = F$. To estimate N^*, we substitute estimates of σ_{ij}^2 and R_j into Eq. 13.

Optimal allocation has proved to be a valuable tool in many sampling situations and should improve the precision of secondary production estimates, even if some of the assumptions, such as the lack of correlation among the $\bar{Y}_{j.}$ values, are violated to a degree. One must be cautious that a design does not leave large "blank spots" that are undersampled or not sampled at all if the system is not well understood. Common sense will often dictate that a sampling scheme somewhere between uniform effort on each date and the indicated optimal allocation regime is most desirable. In view of the tedious effort usually involved in processing invertebrate samples, a little time spent performing optimal allocation calculations should be well justified, considering the possible gains in precision and the reduction in study effort.

Dennis M. Heisey[1]

Minnesota Department of
Natural Resources
Section of Wildlife
500 Lafayette Road
St. Paul 55146

John M. Hoenig[2]

Minnesota Department of
Natural Resources
Section of Fisheries

References

BENKE, A. C. 1979. A modification of the Hynes method for estimating secondary production with particular significance for multivoltine populations. Limnol. Oceanogr. **24:** 168–171.

COCHRAN, W. G. 1977. Sampling techniques. Wiley.

KISH, L. 1965. Survey sampling. Wiley.

KRUEGER, C. C., AND F. B. MARTIN. 1980. Computation of confidence intervals for the size-frequency (Hynes) method of estimating secondary production. Limnol. Oceanogr. **25:** 773–777.

NEWMAN, R. M., AND F. B. MARTIN. 1983. Estimation of fish production rates and associated variances. Can. J. Fish. Aquat. Sci. **40:** 1729–1736.

PERRY, J. N. 1981. Taylor's power law for dependence of variance on mean in animal populations. Appl. Stat. **30:** 254–263.

RESH, V. H. 1979. Sampling variability, life history features, and the experimental design of aquatic insect studies. J. Fish. Res. Bd. Can. **36:** 290–311.

———, AND D. G. PRICE. 1984. Sequential sampling: A cost-effective approach for monitoring benthic macroinvertebrates in environmental impact assessments. Environ. Manage. **8:** 75–80.

WATERS, T. F. 1979. Influence of benthos life history upon the estimation of secondary production. J. Fish. Res. Bd. Can. **36:** 1425–1430.

Submitted: 7 June 1984
Accepted: 29 July 1985

[1] Present address: NH Analytical Software, 801 W. Iowa Ave., St. Paul, Minn. 55117.

[2] Present address: Martin-Marietta Environmental Systems, 9200 Rumsey Road, Columbia, Md. 21045-1934.