from Modeling and Management of Resources under Uncertainty (T. Vincent et. al., eds). Lecture Notes in Biomathematics 72 (1987)

# SAMPLING HIGHLY AGGREGATED POPULATIONS WITH APPLICATION TO CALIFORNIA SARDINE MANAGEMENT

Dedicated to the memory of
Philip Morse:
An extraordinary man and scientist
who pioneered bringing the approach
of basic science to operational problems

Marc Mangel
Departments of Agricultural Economics,
Entomology and Mathematics*
University of California
Davis, California 95616

The management of Pacific sardine off the California coast is used to motivate a number of problems associated with egg or larvae sampling. The use of the negative binomial distribution as a model for the spatial distribution of eggs is discussed and inference for the negative binomial by classical and Bayesian methods is introduced. A particular problem, that of presence-absence sampling when not all sampling sites are habitats, is analyzed in detail. The paper closes with a number of open questions, ranging from improvements in the modeling, to prescriptive problems associated with survey design.

## INTRODUCTION

The work discussed in this paper was motivated by problems associated with the management of the Pacific sardine (*Sardinops Sagax*) in and near the California current. This particular stock - immortalized by John Steinbeck's Doc, Mac and the boys - is estimated to have peaked at a spawning biomass of more than 11,000,000 metric tons and during the cannery heydays (say 1900-1935) fluctuated between about 2,000,000 metric tons and 9,200,000 metric tons (see Smith, 1978 for details about these estimates). By 1965, the spawning biomass had dropped to less than 10,000 metric tons. Currently, state law requires the California Department of Fish and Game to determine on an annual basis whether or not the spawning biomass exceeds 20,000 short tons (1 short ton = .907 metric tons); if it does, then a modest fishery for sardine may be opened. The problem of estimating such a small biomass by standard methods is a very thorny one (see McCall, 1984a,b; Wolf, 1985 for details) and most of the existing methods simply will not work with any accuracy. Consequently, Wolf and Smith (1985) proposed an "inverse egg production method" (IEPM) for determining the spawning biomass. This method is based on the idea that as the spawning biomass increases, the area in which eggs are found will increase.

*Address correspondence to the Department of Mathematics.

Operationally, the method proceeds as-follows. One lays down a sampling pattern, as shown in Figure 1. Each dot in Figure 1 represents a station at which eggs are sampled. The region shown in Figure 1 is about five times larger than areas occupied by the Pacific sardine in recent years. Stations are 10 nm (nm = nautical miles; 1 nm = 1853 meters) apart going NW to SE and 4 nm going offshore so that each station represents 40 $(nm)^2$. The idea behind the IEPM is to estimate a priori the area that a 20,000 short ton spawnings biomass would occupy, then to sample for eggs and determine if this critical area $A_c$ is exceeded. If it is, then it is likely that the spawning biomass exceeds 20,000 short tons. When this method was applied in 1985, 419 stations were sampled. Eleven of these stations had eggs; about 85 eggs were discovered. On the basis of these data and IEPM, a 1000 ton sardine fishery was recommended in 1985 (P. Smith, NMFS, La Jolla, personal communication).

This paper is concerned with various modelling and analytical issues associated with egg surveys similar to the one just described. (These kinds of problems, however, are broader than egg surveys - see Downing (1979) or Resh (1979) for other kinds of applications and motivations.)

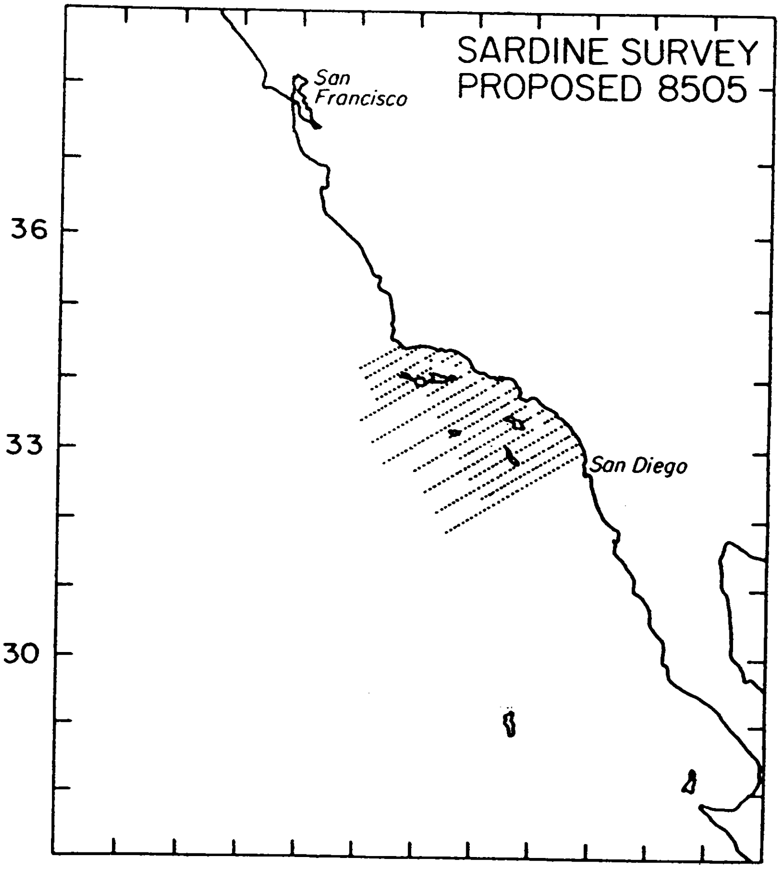

Figure 1. The sampling sites for the 1985 sardine egg survey proposed by NMFS Scientists (taken from Wolf and Smith (1984)).

Some of these issues are the following: what kinds of models should be used for highly aggregated populations and why; how does one extract the maximum information; how does one deal with data that involve a preponderance of zero's in the samples? In the spirit of a workshop paper, these issues are discussed from the current viewpoint of the author (i.e., subject to possible change) and various untested ideas are presented as a way to probe their usefulness.

## MODELING IDEAS AND ISSUES

This section contains a discussion of a number of pertinent questions associated with modeling sampling surveys for highly aggregated populations. To begin, one should think about the spatial scales of interest. These are

| Entity | Spatial Scale |
|---|---|
| individual fish | ~ cm |
| school | ~ 100 m |
| egg patches | ~ 1000 m |
| school groups | ~ 10000 m |
| sampling scale | ~ 10000 m |

Thus, the sampling scale is large enough to justify the assumption that numbers of eggs taken at different stations are independent random variables. Let $X_i$ be a random variable representing the number of eggs taken in the sample at the $i^{th}$ station, which will henceforth be called a site. Some of the properties of $X_i$ should be the following ones:

$$\Pr\{X_i = 0\} \text{ should be considerable} \tag{1}$$

$$V\{X_i\} >> E\{X_i\} \tag{2}$$

where $V\{X_i\}$ is the variance of $X_i$ and $E\{X_i\}$ is the expected value (mean of $X_i$). Properties of (1) and (2) are based on the experimental reality, not any theoretical conceptualization.

For most of this paper, the following model is used: If the $i^{th}$ site is a habitat, the conditional distribution of $X_i$ is a negative binomial with parameters m and k (written NB(m,k)). That is

$$\Pr\{X_i = x \mid \text{site is a habitat}\} = \frac{\Gamma(k+x)}{\Gamma(k)x!} \left(\frac{k}{k+m}\right)^k \left(\frac{m}{k+m}\right)^x \tag{3}$$

where $\Gamma(\cdot)$ is the gamma function. The distribution (3) has the properties (1) and (2) (see Johnson and Kotz (1969) for a discussion of more properties of the NB(m,k) distribution). First

$$\Pr\{X_i = 0\} = \left(\frac{k}{k+m}\right)^k \tag{4}$$

which can be considerable even if m is large (see Figure 2). Second

$$\left.\begin{aligned} E\{X_i\} &= m \\ V\{X_i\} &= m + \frac{m^2}{k} \end{aligned}\right\} \tag{5}$$

so that if k is small, property (2) is satisfied. What value of k should be used? Smith and Richardson (1977) provide the following data

| | Spawning Biomass (Millions of Tons) | Estimate of k |
|---|---|---|
| | 3.9 | .14 |
| | 2.7 | .19 |
| | 1.0 | .21 |
| | .2 | .08 |
| Average | 2.95 | .155 |
| Coefficient of Variation | 1.48 | .65 |

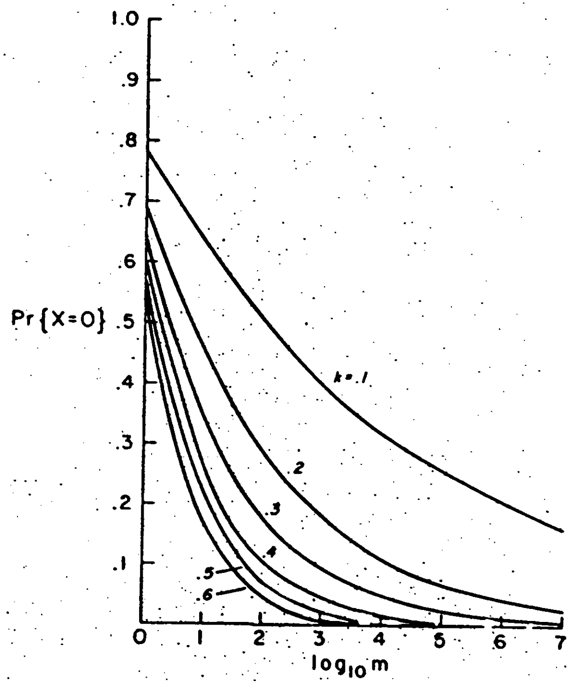

Figure 2. Likelihood of a zero observation in a NB distribution with parameters m and k.

As the spawning biomass varies over a range of 19.5, note that k varies by a factor of 2.6. For all intents and purposes - and definitely for this paper - k will be treated as a constant, presumed known, in the range of 0.1 to 0.2.

One can legitimately ask if there is a true biological or operational motivation for choosing $X_i \sim NB(m,k)$. Here is one. Let B(t) denote the spawning biomass at time t and assume that it satisfies the following stochastic differential equation:

$$dB = B(t)\{r(1-B(t)/K)dt + \sigma dW\} \tag{6}$$

where $dB = B(t+dt) - B(t)$, r, K and $\sigma$ are parameters and $dW = W(t+dt) - W(t)$, where W(t) is Brownian motion (see, e.g., Ludwig, 1975; Schuss, 1980, for general discussions of Brownian motion or Mangel, 1985a for a discussion related to natural resource models). If B(t) satisfies (6), then the equilibrium density for B is a gamma density (see Dennis and Patil, 1984, for an elaboration). Assume that, given a value of the equilibrium biomass $B_{eq}$, the distribution of eggs encountered is a Poisson with parameter $\lambda = \lambda_0 B_{eq}$. Then the unconditional density for the number of eggs is a negative binomial (see, e.g., Mangel, 1985b, for details).

One can also ask if there is a legitimate biological reason for choosing constant k. The answer is, to some extent, yes. In order for eggs to be fertilized, they need to be highly clumped - regardless of the size of the spawning biomass. This will partially justify the use of constant k.

In the analysis which follows in the next section, it is assumed that $X_i$ has the distribution (3) with known k but unknown m and it is assumed that if the spawning biomass exceeds a critical value, then m will exceed a given critical value, $m_c$.

Before presenting this analysis, however, a number of points need to be cleared up. First, the NB(m,k) model is not the only one with properties (1) and (2). For example, one can use other "contagious" distributions such as the Neyman Type A in which

$$\Pr\{X_i = x\} = \sum_{j=1}^{\infty} \frac{e^{-\lambda}\lambda^j}{j!} e^{-j\phi} \frac{(j\phi)^x}{x!} \tag{7}$$

and $\lambda$ and $\phi$ are parameters (see Johnson and Kotz, 1969, Chapter 9). In this model

$$\left.\begin{aligned} \Pr\{X_i = 0\} &= e^{-\lambda(1-e^{-\phi})} \\ E\{X_i\} &= \lambda\phi \\ V\{X_i\} &= \lambda\phi(1+\phi) \end{aligned}\right\} \tag{8}$$

so that the properties (1) and (2) can be satisfied. The NB(m,k) model is used here, but others are feasible.

Second, note that (3) is conditioned on a site being a habitat. Thus, one needs to append to (3)

$$p_i = \Pr\{i^{th} \text{ site is a habitat}\} . \tag{9}$$

Some choices for $p_i$ are

$$p_i = \begin{cases} p_0\text{, a constant} \\ p(m)\text{, a function of } m \\ p(i)\text{, a function of site location} \\ p(i,m)\text{, a function of site location and } m \end{cases} \tag{10}$$

For example, one could use

$$p(m) = 1 - e^{\gamma m} \tag{11}$$

where $\gamma$ is a constant.

Using (9) and (3) leads to the following model

$$\Pr\{X_i = 0\} = (1 - p_i) + p_i \left[\frac{k}{k+m}\right]^k \tag{12}$$

$$\Pr\{X_i > 0\} = p_i \left[1 - (\frac{k}{k+m})^k\right] \tag{13}$$

This will be the basic model in the next section. Note that, in this model, one still assumes that if eggs are present, they're found. But the cell size is 40 nm$^2$ and the sampler size is roughly ~ .05 m$^2$ so that even if eggs are present, they could be missed. A way around this problem is discussed in the last section.

Third, one should separate descriptive and prescriptive sampling problems. The prescriptive problems are "survey optimization" ones: how should optimal surveys be connected? These are sexy problems, but often of less use to managers than the descriptive problem of "here's the data, what does it mean?" In the next section, a descriptive problem and its analysis are described. Optimal surveys are discussed in the last section.

## PRESENCE-ABSENCE SAMPLING FOR EGGS

In this section, the ideas developed thus far are applied to a problem which is analogous to the sardine egg sampling problem (the details of where the analogy fails are discussed in the next section). The set-up for the problem is this: the $i^{th}$ site is a habitat with probability $p_i$ and one samples for the presence or absence of eggs (ignoring the actual number of eggs, if there are eggs present). This scheme, in operational terms, would allow the survey scientist to bring the samples back in, hold it up to the light and determine the presence or absence of eggs.

As before, $X_i$ is the number of eggs in the sample at the $i^{th}$ site and $X_i = 0$ is called a "negative" sample; $X_i > 0$ is called a positive sample. Equations (12) and (13) give the probabilities that $X_i = 0$ and $X_i > 0$ respectively.

Assume that there are $N_n$ negative samples and N positive samples. Let $n$ denote those sites at which a negative sample was obtained and p denote those sites at which a positive sample was obtained. The likelihood of {n,p} is

$$\mathcal{L}(n,p|m) = \prod_{i\epsilon n} \{1-p_i+p_i\ (\frac{k}{k+m})^k\} \prod_{i\epsilon p} \{p_i-p_i\ (\frac{k}{k+m})^k\} \qquad (14)$$

A number of different kinds of sampling schemes can be derived, based on assumptions about the values of $p_i$. Some of these will now be discussed.

First consider the case in which all the $p_i$ take the same value, p. Then (14) becomes

$$\mathcal{L}(N_n,N|m) = \left[1-p+p\ (\frac{k}{k+m})^k\right]^{N_n} \left[p-p\ (\frac{k}{k+m})^k\right]^N \qquad (15)$$

Note the following about (15): the model is now essentially a binomial model, with success probability $p-p(k/(k+m))^k$. Thus, the likelihood in (15) is the unnormalized probability of N successes in $N_n$ + N trials. The normalization constant, a binomial coefficient, is not needed for any of the calculations that follow.

The maximum likelihood value of $m,\hat{m}$ is found by taking the derivative of the logarithm of $\mathcal{L}(N_n, N|m)$ with respect to m and setting it equal to 0. This leads to a nonlinear equation for m, which is easily solved on a desktop microcomputer. Once the MLE $\hat{m}$ is known, one can investigate likelihood ratios for other values of m. This approach will be reported elsewhere (Mangel and Smith, 1987).

Instead, consider a Bayesian approach in which one wishes to compute the posterior probability that $m > m_c$, given the data. In order to do this, one needs a prior distribution of m. Two choices are the uniform prior

$$f_0(m) = 1 \qquad 0 < m < m_m \qquad (16)$$

where $m_m$ is a specified value, and the noninformative prior

$$f_0(m) = \frac{1}{\sqrt{m(k+m)}} \qquad (17)$$

(The noninformative prior (17) is derived in the appendix.)

These are chosen to represent "ignorance" about the value of m. When the uniform prior (UP) is used, all values of m between 0 and $m_m$ are given equal weighting. When the noninformative prior (NP) is used, data change the position, but not the shape of the posteriori distribution (see Martz and Waller, 1982, for further discussion).

If the uniform prior is used, the posterior probability that m exceeds $m_c$ is given by

$$P_{UP}(m > m_c) = \frac{\int_{m_c}^{m_m} \left[1-p+p(\frac{k}{k+m})^k\right]^{N_n}\left[p-p(\frac{k}{k+m})^k\right]^N dm}{\int_0^{m_m} \left[1-p+p(\frac{k}{k+m})^k\right]^{N_n}\left[p-p(\frac{k}{k+m})^k\right]^N dm} \qquad (18)$$

Since $m_m$ may be quite large (say of the order of 1000) it helps to introduce

$$\left.\begin{aligned} w &= \frac{k}{k+m} \\ dw &= -\frac{k}{(k+m)^2}\, dm = -\frac{w^2}{k}\, dm \\ w_c &= \frac{k}{k+m_c} \qquad w_m = \frac{k}{k+m_m} \quad . \end{aligned}\right\} \qquad (19)$$

The integral in (18) becomes

$$P_{UP}(m > m_c) = \frac{\int_{w_m}^{w_c} \frac{[1-p+pw^k]^{N_p}[p-pw^k]^N dw}{w^2}}{\int_{w_m}^{1} \frac{[1-p+pw^k]^{N_p}[p-pw^k]^N dw}{w^2}} \quad . \qquad (20)$$

These integrals are easily computed on a desktop microcomputer.

When the noninformative prior is used, one makes the transformation

$$\left.\begin{aligned} m &= k^2 \tan\theta \\ dm &= 2k \tan\theta \; d\theta/\cos^2\theta \\ k + m &= k/\cos^2\theta \end{aligned}\right\} \qquad (21)$$

$$\theta_c = \text{arc tan} \left(\sqrt{\frac{m_c}{k}}\right)$$

and finds that the posterior probability is given by

$$P_{NP}(m > m_c) = \frac{\int_{\theta_c}^{\theta} [1-p+p(\cos\,\theta)^{2k}]^{N_n}[p-p(\cos\,\theta)^{2k}]^N \frac{d\theta}{\cos\,\theta^2}}{\int_0^{\theta_m} [1-p+p(\cos\,\theta)^{2k}]^{N_n}[p-p(\cos\,\theta)^{2k}]^N \frac{d\theta}{\cos\,\theta^2}} \qquad (22)$$

Figure 3 shows $P\{m > m_c\}$ as a function of the number of positive samples using both priors. The NP is more "conservative" than the UP.

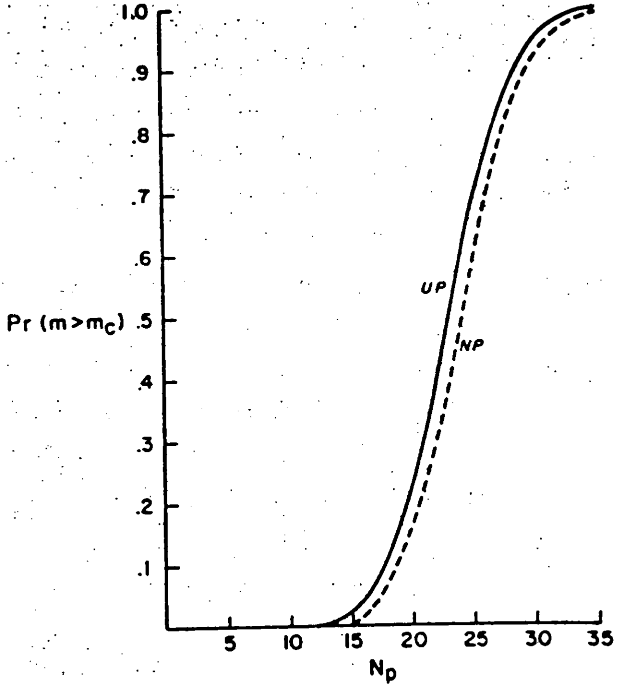

Figure 3. Probability that m exceeds $m_c$ = 1.14 as a function of the number of positive samples (N) in a total of N = 100 samples. Other parameters: $m_m$ = 1000, k = .2, p = .8.

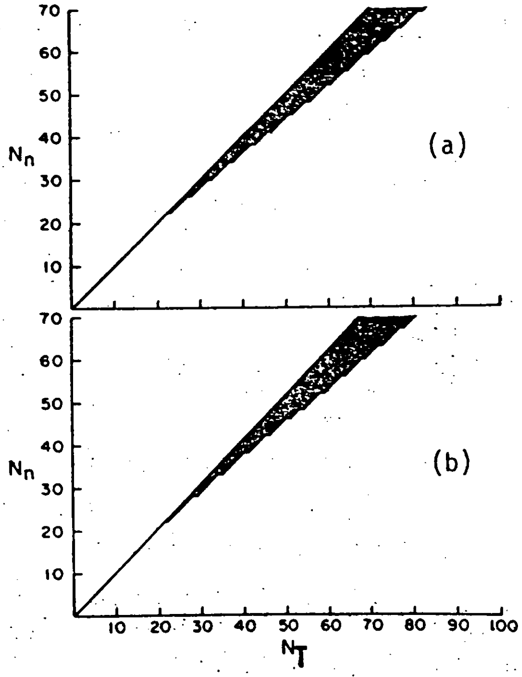

Figure 4. Sequential sampling charts in which the number of negative samples (N) is plotted against the total number of samples $N_T$. If the data fall in the shaded region, one can conclude with 99% confidence (Figure 4a) or 95% confidence (Figure 4b) that m < $m_c$. Other parameters: $m_c$ = 1.14, k = .2, p = .8, $m_m$ = 1000. The uniform prior was used in the calculations.

Figure 3 is an ex post facto probability statement made after the data are collected. On the other hand, for many situations a sequential sampling plan is often more useful. Figure 4 is a sequential sampling diagram used to compute the probability that $m < m_c$ under the uniform prior. In this diagram, one plots $N_n$ versus $N_T = N + N_n$. If an observation falls in the shaded region, then one can conclude that $m < m_c$ with probability .99 (Fig. 4a) or .95 (Fig. 4b). If the current data point $(N_T, N_n)$ does not fall in the shaded region, then an additional site is sampled.

These same kinds of calculations can be performed when the generalized likelihood (14) is used. For example, when the UP is used, one finds

$$P_{UP}(m < m_c) = \frac{\int_{w_c}^{1} \prod_{i \epsilon n} \{1-p_i+p_i w^k\} \prod_{i \epsilon p} (p_i-p_i w^k) \frac{dw}{w^2}}{\int_{w_m}^{w_c} \prod_{i \epsilon n} \{1-p_i+p_i w^k\} \prod_{i \epsilon p} (p_i-p_i w^k) \frac{dw}{w^2}} \qquad (23)$$

The only difficulty is that one cannot develop charts similar to Figures 3 and 4. On the other hand, (23) is ideal for use in real-time with a microcomputer. For example, assume that $m_c = 1.14$, $m_m = 1000$, $k = .2$ and let each data point $(p_i, X_i)$ be represented with $X_i = 1$ for a positive sample and $X_i = 0$ for a negative sample. Suppose that the first 10 data points are (1,0), (1,0), (.95,0), (.95,0), (.9,0), (.9,0), (.85,0), (.85,0), (.8,0), and (.8,0). Using (23) shows that $P_{UP}(m < m_c) = .82$. If the next five data points are (1,0), (1,0), (.95,0), (.95,0), and (.9,1), then $P_{UP}(m < m_c) = .86$. If the next five data points are (1,0), (1,0), (.95,0), (.95,0) and (.9,0), then $P_{UP}(m < m_c) = .96$ and sampling can stop if a 95% confidence level is desired.

Two points are worth noting. First, there is a preponderance of zeroes in the data. This kind of result is, in fact, observed in sampling. Second, a large amount of negative information is needed to insure that $m < m_c$ with a high confidence level. One must remember, however, that with the UP, the initial probability that $m < m_c$ is $m_c/m_m$. So, for example, for the values presented here, the prior probability that $m < m_c$ is $1.1 \times 10^{-3}$. In addition, since not every site is a habitat, the effects of negative information on the updated distribution are mitigated (i.e., as $P_i \rightarrow 0$, the data have decreasing effects on the Bayesian update).

## OPEN QUESTIONS

Since a major purpose of a workshop is to raise questions, a workshop paper can (and perhaps should) end with open questions rather than conclusions. In this spirit, a number of open questions and directions for future work are indicated.

1) Many Age Classes of Eggs. In the actual sardine survey, four age classes of eggs (<1 day, 1-2 days, 2-3 days old and >3 days old) are sampled, each with a different m and k. Thus, the data are more complicated, consisting of presence-absence of the four age classes or the actual counts of the four age classes. The question of how to use these data is complex. One could assume, for example, that the four age classes represent completely independent events (probably an unrealistic assumption). The other extreme is one of complete correlation: if any age class is present, then they all are. Reality probably lies somewhere between the two extremes, with a partial correlation. A Bayesian approach to this problem can also be developed (see Mangel et al., 1984, pg. 568) where the correlation level is a user-inputted variable.

For example, let $m_j$ and $k_j$ denote the unknown mean and known aggregation parameter for the $j^{th}$ egg class. A reasonable model is the following one (P. Smith, NMFS, La Jolla, personal communication):

$$\left.\begin{array}{ll} k_j = .1j & j = 1,2,3,4 \\ m_j = s_j m \text{ where m is unknown and} & \\ s_1 = 1 & \\ s_2 = .8 & \\ s_3 = .6 & \\ s_4 = .4 & \end{array}\right\} \qquad (24)$$

Finally, let $\rho_c$ denote a correlation parameter, in the following sense: with probability $\rho_c$, if eggs of one age class are present, then eggs of all age classes are present. With this kind of model, the probability of a negative sample is

$$\Pr\{X_i=0\} = (1-\rho_c) \prod_{j=1}^{4} \left(\frac{k_j}{k_j+s_j m}\right)^{k_j} + \rho_c \max_j \left(\frac{k_j}{k_j+s_j m}\right)^{k_j} \qquad (25)$$

One can do similar sorts of analyses using (25). Preliminary investigations based on (25) with four age classes of eggs (as in (24)) support the recommendation of opening a small sardine fishery in 1985.

2) Imperfect Sampling. Another possible extension allows for the chance of imperfect sampling. One way to do this is to use the weighted NB(WNB) model of Bissell (1972). According to that model, if a site is a habitat

$$\Pr\{X_i=x\} = \frac{\Gamma(k+x)}{x!\,\Gamma(k)} \left(\frac{k}{mW_i+k}\right)^{k} \left(\frac{mW_i}{mW_i+k}\right)^{x} \qquad (26)$$

where $W_i$ is a measure of sampling efficiency. (Zweifel and Smith (1981) discuss the interpretation of $W_i$.) The data now consist of triplets $(p_i, W_i, X_i)$. The methods of the previous sections can be extended to cover this case with essentially no conceptual difficulty and only minor computational difficulty.

Another way to do imperfect sampling is to take into account explicitly the chance that an egg patch might be missed during the samplings. For example, let $B_i$ denote the number of eggs in the cell containing the ith site and let A(b) denote the area of patch with b eggs. One choice is

$$A(b) = A_m(1 - e^{\gamma b}) \tag{27}$$

where $A_m$ and $\gamma$ are constants. One expects $A_m \ll S$ where S is the 40 $nm^2$ area of each cell. Finally, assume that

$$\Pr\{\text{detecting eggs} \mid B_i = b\} = \frac{A(b)}{S} \tag{28}$$

There are now two ways for $X_i = 0$: no eggs present ($B_i = 0$) or eggs present, but missed. Thus,

$$\Pr\{X_i=0\} = \Pr\{B_i=0\} + \sum_{B=1}^{\infty} \Pr\{B_i=b\} \Pr(\text{no detection} \mid B_i=b\}$$

$$= \left(\frac{k}{k+m}\right)^k + \sum_{b=1}^{\infty} \Pr\{B_i=b\} \left[1 - \frac{A_m}{S}(1-e^{-\gamma b})\right] \tag{29}$$

After a little algebra, and use of the generating function for a NB distribution, one can develop an explicit formula for $\Pr\{X_i=0\}$ in terms of m, k, $A_m$, S and $\gamma$. The development and use of this formula will be given in Mangel and Smith (1987).

3) Joint Estimation of Habitat Boundaries and m. An open question, which may require a new formulation of the problem, involves the simultaneous estimation of the habitat boundary and m. That is, one might consider the joint density that $p_i = 0$ and m takes a certain value. The approach to this problem is not clear, although a Bayesian formulation seems natural.

4) Ideal-Free Sardine Eggs. At the workshop, Mike Rosenzweig pointed out the ecological theory of habitat selection could be used to generate stratified sampling plans. That is, use habitat theory (see Rosenzweig's article in this volume) to predict the proportion of sites with eggs in habitats of different quality as a function of spawning biomass. If one could identify habitat quality on the basis of oceanographic factors (e.g., satellite photographs of temperature and chlorophyll distributions), then it would be possible to use a sampling scheme based on habitat

quality. Such a scheme might require considerably less information (i.e., fewer samples).

5) Adaptive Survey Optimization. Of all the possible prescriptive problems, perhaps the most interesting one is the development of an adaptive (i.e., closed loop) algorithm which can be used to guide the survey vessel. That is, based on the sampling history thus far, which site should be visited next.

6) Economic Modeling. Recall that the purpose of the egg survey is to determine a level of confidence about the biomass and that if the biomass exceeds a critical level, then a complete stock survey will be conducted. One can extend the methods of this paper to include the costs of the egg survey, the cost of the complete stock survey, and the cost of not allowing fishing when the stock exceeds the critical level.

7) Egg Surveys as Priors. Assuming that one decides to pursue a complete stock survey. The results of the egg survey can be used as a prior density when planning the larger survey. The results presented in the previous section on estimating the extent of the habitat could be especially useful.

## ACKNOWLEDGMENTS

This work was partially supported by NSF Grant MCS-81-21659, by the Agricultural Experiment Station of the University of California, and by NOAA through the California Sea Grant Project. I thank Richard Plant, Patti Wolf, Don Ludwig and Carl Walters for useful discussions and Paul Smith for his willingness to share ideas, time and data.

## REFERENCES

Anscombe, F.J. 1950. Sampling theory of negative binomial and logarithmic series distributions. Biometrika, Vol. 34, pp. 358-382.

Bissell, A.F. 1972. A negative binomial model with varying element sizes. Biometrika, Vol. 59, pp. 435-441.

Bliss, C.I. 1958. The analysis of insect counts as negative binomial distributions. Proc. Tenth Intl. Cong. Entom., pp. 1015-1032.

Box, G.E.P. and G.C. Tiao. 1973. Bayesian Inference in Statistical Analysis. Addison Wesley, Reading, MA, 588 pp.

DeGroot, M. 1970. Optimal Statistical Decisions. McGraw-Hill, NY, 489 pp.

Dennis, B. and G.P. Patil. 1984. The gamma distribution and weighted multimodal gamma distributions as models of population abundance. Mathematical Biosciences, Vol. 68, pp. 187-212.

Downing, J.A. 1979. Aggregation, transformation and the design of benthos sampling programs. J. Fisheries Res. Board of Can., Vol. 36, pp. 1454-1463.

Feller, W. 1968. An Introduction to Probability Theory and its Applications, Vol. 1, John Wiley, NY, 509 pp.

Gerard, G. and P. Berthet. 1971. Sampling strategy in censusing patchy populations. In G.P. Patil, E.C. Pielou and W.E. Waters (eds.), Statistical Ecology, Vol. 1, pp. 59-68, Pennsylvania State University Press, University Park, PA.

Gunderson, D.R., G.L. Thomas, P. Cullenberg, D.M. Eggers and R.F. Thorne. 1980. Rockfish investigations off the coast of Washington. Report FRI-UW-8021, Fisheries Research Institute, University of Washington, Seattle.

Hewitt, R. 1976. Sonar mapping in the California Current area: A review of recent developments. Cal. COFI Report, Vol. 18, pp. 149-154.

Hewitt, R. 1981. The value of pattern in the distribution of young fish. Rapp. P-v. Reun. Cons. Int. Explor. Mer., Vol. 178, pp. 229-236.

Hewitt, R. 1984. 1984 Spawning biomass of Northern Anchovy. Administrative Report LJ-84-18, Southwest Fisheries Center, National Marine Fisheries Service, La Jolla, CA.

Hewitt, R. and P. Smith. 1979. Seasonal distributions of epipelagic fish schools and fish biomass over portions of the California Current region. Cal. COFI Report, Vol. 20, pp. 102-110.

Hewitt, R., P.E. Smith, and J.C. Brown. 1976. Development and use of sonar mapping for pelagic stock assessment in the California Current area. Fish Bull. US, Vol. 74, pp. 281-300.

Hewitt, R. and P.E. Smith. 1982. Sonar mapping of the California Current area: Some considerations of sampling strategy. Report, Southwest Fisheries Center.

Johnson, N. and S. Kotz. 1969. Discrete Distributions in Statistics. Wiley, NY.

Leaman, B.M. 1981. A brief review of survey methodology with regard to groundfish stock assessment. Can. Spec. Pub. Fish. Aq. Sci., Vol. 58, pp. 113-123.

Lloyd, M. 1967. Mean crowding. J. Anim. Ecol., Vol. 36, pp. 1-30.

Ludwig, D. 1975. Persistence of dynamical systems under random perturbations. SIAM Review, Vol. 17, pp. 605-640.

MacCall, A.D. (ed.) 1984a. Report on a NMFS-CDFG workshop on estimating pelagic fish abundance. Administrative Report LJ-84-40, Southwest Fisheries Center, POB 271, La Jolla, CA 92038.

MacCall, A.D. (ed.) 1984b. Management information document for California coastal pelagic fishes. Southwest Fisheries Center Administrative Report LJ-84-39. Southwest Fisheries Center, POB 271, La Jolla, CA 92038.

Mangel, M. 1985a. Decision and Control in Uncertain Resource Systems. Academic Press, NY.

Mangel, M. 1985b. Search models in fisheries and agriculture. In M. Mangel, (ed.), Proc. of the Ralf Yorque Workshop on Resource Management, Springer Verlag, NY.

Mangel, M. and P.E. Smith. 1987. Presence-absence plankton sampling for fisheries management. Can. J. Fish. Aq. Sci., to appear.

Martz, H. and R. Waller. 1982. Bayesian Reliability Analysis. John Wiley and Sons, NY. 745 pp.

Pennington, M. 1983. Efficient estimators of abundance, for fish and plankton surveys. Biometrics, Vol. 39, pp. 281-286.

Pennington, M. and P. Berrien. 1984. Measuring the precision of estimates of total egg production based on plankton surveys. J. Plankton Res., Vol. 6, No. 5, pp. 869-880.

Pielou, E.C. 1977. Mathematical Ecology. Wiley, NY. 385 pp.

Resh, V.H. 1979. Sampling variability and life history features: Basic considerations in the design of aquatic insect studies. J. Fish. Res. Board Can., Vol. 36, pp. 290-311.

Schuss, Z. 1980. Theory and Application of Stochastic Differential Equations. Wiley, NY.

Smith, P.E. 1978. Biological effects of ocean variability: Time and space scales of biological response. Rapp. P-v. Reun. Cons. Int. Explor. Mer., Vol. 173, pp. 117-127.

Smith, P.E. and S.L. Richardson. 1977. Standard techniques for pelagic fish egg and larva surveys. FAO Fisheries Technical Paper 175, Food and Agriculture Organization of the United Nations, Rome, Italy.

Taylor, C.C. 1953. Nature of variability in trawl catches. Fish. Bull. 83, U.S. Department of the Interior, Vol. 54, pp. 145-166.

Taylor, L.R. 1971. Aggregation as a species characteristic. In G.P. Patil, E.C. Pielou and W.E. Waters (eds.), Statistical Ecology, Vol. 1, pp. 357-377, Pennsylvania State University Press, University Park, PA.

Wald, A. 1947. Sequential Analysis. Dover, NY. 121 pp.

Wolf, P. 1985. Status of the spawning biomass of the Pacific Sardine, 1984-85. Marine Resources Report to the Legislature, California Department of Fish and Game.

Wolf, P. and P.E. Smith. 1985. An inverse egg production method for determining the relative magnitude of Pacific sardine spawning biomass off California. Cal. COFI Report, Vol. 26, pp. 130-138.

Zweifel, J.R. and P.E. Smith. 1981. Estimates of abundance and mortality of larval anchovies (1951-75): Application of a new method. Rapp. P-v. Cons. Int. Explor. Mer., Vol. 178, pp. 248-259.

## APPENDIX: DERIVATION OF THE NONINFORMATIVE PRIOR

The approximate noninformative prior for the NB distribution is derived as described by Martz and Waller (1982, pg. 224). Viewing (3) as the likelihood of m given x, the log-likelihood is

$$L(m|x) = -k \log(k+m) + x[\log m - \log(m+k)] + \ell(x,k) \qquad \text{(A-1)}$$

where $\ell(x,k)$ contains terms independent of m. The derivatives of the log-likelihood are

$$\frac{\partial L}{\partial m} = -\frac{k}{k+m} + \frac{x}{m} - \frac{x}{m+k}$$

$$\frac{\partial^2 L}{\partial m^2} = \frac{k}{(k+m)^2} - \frac{x}{m^2} + \frac{x}{(m+k)^2} \tag{A-2}$$

Setting $\partial L/\partial m = 0$ shows that the maximum likelihood estimate is $\hat{m} = x$. (For n independent observations, the MLE $\hat{m}$ is easily shown to be the sample mean.) Define

$$J(\hat{m}) = -\frac{\partial^2 L}{\partial m^2}\Big|\hat{m} = \frac{\hat{m}}{\hat{m}^2} - \Big/ \frac{\hat{m}+k}{(k+\hat{m})^2} = \frac{k}{\hat{m}(k+\hat{m})} \tag{A-3}$$

The approximate non-informative prior is then

$$f_0(m) \propto J(m)^{1/2} \propto m^{-1/2}\,(k+m)^{-1/2} \tag{A-4}$$

## PARTICIPANT'S COMMENTS

Mangel's paper addresses a relatively common problem experienced by field biologists, i.e., the non-random distribution of organisms in nature. The approach he is suggesting, use of discrete distribution statistics rather than strict reliance upon the Central Limit Theorem and large numbers of samples to yield good approximations to the Gausian distribution, is likely to become increasingly popular as microcomputers become standard field equipment. The negative binomial distribution has enjoyed some attention among biologists in the past (see Elliott (1977) for examples), but the Pacific sardine example is a particularly interesting case because the presence or absence of eggs in the spawning ground survey determines whether or not the fishery will open in a given year. The 1985 results, 85 eggs at 11 stations, led to the very surprising recommendation to open a 1000 ton sardine fishery! This appears to be very little empirical information upon which to base such a decision, and it immediately raises other technical questions such as the certainty of identification of the eggs, or determination of their ages.

In the Open Questions section, Mangel raises several interesting possibilities for further development of the model. If the negative binomial is really a good model for the underlying distribution of eggs, then the k parameter may be a useful index of dispersion. It is sensitive to size and number of sampling units but within a particular survey it should be useful as a relative index. The four different age classes of eggs (each with different m and k) should show a progression of dispersion with time since spawning. This may allow inference about the number of spawning aggregations present in the region. This hypothesis could be tested with historical data if survey information is available from periods in which there has been some variation in abundance of the spawning stock.

Further consideration of sampling methods appropriate to contagious distributions is long past due in fisheries management. The problem of estimating angler effort in recreational fishing surveys (creel surveys) may be amenable to sampling from satellite imagery where the boats or ice-houses concentrate over known habitats or, more directly, on schools of fish. The degree of aggregation of fishing units may prove to be a useful index of habitat quality or a covariate with the catchability coefficient. Improvements in our ability to estimate the latter statistic would greatly enhance the utility of most of the catch-effort models currently used in fishery management.

## Reference

Elliott, J.M. 1977. Some methods for the statistical analysis of samples of benthic invertebrates. Sci. Publ. No. 25, Freshwater Biological Association, Ambleside, Cumbria: 160 pp.

George R. Spangler

The Mangel method promises a revolution in management techniques. As I have indicated in my own chapter of the workshop, it may be possible to refine it by consciously sorting the sampling sites according to habitat quality, and then sampling them in inverse order of their quality (poorest first, etc.).

It is not often that good management practices can actually improve the theory that produces them, but this time that may happen. Practicing the Mangel method will require developing a habitat quality index, $\underline{Q(\underset{\sim}{h})}$, for the species being managed (where $\underset{\sim}{h}$ is a vector of habitat properties). It also requires knowing at which population density each quality of habitat is added. Let us standardize the densities by dividing them by a constant obtained through sampling (say $\bar{Y}$, the average yield per unit fishing effort). Then the management data will determine the function:

$$Q(\underset{\sim}{h})^* = f_i(\underline{N_i/\bar{Y}_i})$$

where $Q(\underset{\sim}{h})^*$ is the habitat quality which is marginally used when the standard density of the $\underline{i}$-th species is $N_i/\bar{Y}$.

The subscript $\underline{i}$ introduces the possibility of a multispecies view, and here is where the need for data collection by managers will enhance the basic science on which they rely. Community ecologists are investigating the fundamental structure of sets of niches (Rosenzweig, in press; P.S. Giller and J.H.R. Gee (eds.), 27th British Ecological Symposium, 1986: Organization of Communities: past and present). Eventually, this knowledge should also benefit the management of exploited populations, but it is still too rudimentary for that. However, obtaining the set of functions, $\underline{f_i}$, for a guild of species on one quality index, $\underline{Q(\underset{\sim}{h})}$, will illuminate what is happening among the species. If, for example, all the $\underline{f_i}$ are positive functions, then the species rank the habitats similarly. If some are negative, then they have at least two distinct habitat preferences. If they are all unimodal, but peak at various places along the $\underline{N/\bar{Y}}$ axis, then they all have distinct preferences. Other patterns, of which no one has yet conceived, may emerge. But there is no question that knowing the peaks of the various species and how they relate to each other will advance community ecology and may suggest improvements in management policies.

Michael L. Rosenzweig