# SAMPLING SURVEYS FOR HIGHLY AGGREGATED POPULATIONS

by

Marc Mangel

Department of Mathematics
University of California
Davis, California 95616

# SAMPLING SURVEYS FOR
# HIGHLY AGGREGATED POPULATIONS

by

Marc Mangel

1986

Report 86-3 on Sea Grant Project 6-F-N
"New Methods in Stock Abundance Estimation"

California Sea Grant
College Program, R/F-109

# Sampling Surveys for Highly Aggregated Populations

Marc Mangel*
Department of Mathematics
University of California
Davis, California 95616

In this paper, a number of sampling problems for populations that have an underlying negative binomial distribution with mean m and aggregation parameter k (i.e., Pr {observing a level x} = $(\Gamma(k+x)/\Gamma(k)x!)(\frac{k}{k+m})^k(\frac{m}{k+m})^x$ ) are discussed. It is assumed that k is known but m is unknown, with a prior distribution $f_0(m)$. The objective is to determine if m exceeds a critical level $m_c$. Three models are introduced: presence absence sampling when all sampling sites are habitats, presence-absence sampling when a site may not be a habitat, and actual updating of the mean. Both fixed sample and sequential procedures are described using a Bayesian approach. The problem of knowing when a habitat has been exited is also discussed. The theory is motivated and applied to egg surveys for Pacific Sardine. In the Appendix, a new model (Zero/Random or Z/R) for aggregation is proposed.

* Also: Aquaculture and Fisheries Program, Departments of Agricultural Economics, Entomology

A key element in biological resource management is having some estimate of the current state of the resource (a specific example - Pacific Sardine - is described in the next section). In order to obtain such information, fishery scientists very often run surveys to sample the resource (e.g., fish schools) or some proxy to it (e.g., egg or larval surveys). This paper is concerned with questions in the design of such surveys when the sampled population is highly aggregated. Aggregation is common property of many biological species (e.g., Taylor (1971), Bliss (1958)) and is certainly observed in populations of fish (see, e.g., Taylor (1953), Gunderson et al. (1980), Smith (1978 a,b), Zweifel and Smith (1981)), Hewitt (1978)). (A biological mechanism for aggregation is discussed in the next section). One way to describe the aggregation is to assume that the population follows a negative binominal (NB) distribution (Pielou (1977), Mangel (1984)). That is, if $X$ is a random variable corresponding to the observation, then

$$\Pr\{X=x\} = \frac{\Gamma(k+x)}{\Gamma(k)x!}\left(1+\frac{m}{k}\right)^{-k}\left(\frac{m}{m+k}\right)^{x} \tag{1}$$

In this equation, $\Gamma(\cdot)$ is the gamma function (for integer values $\Gamma(n) = (n-1)!$) and $m$ and $k$ are parameters. In particular (Feller, 1971)

$$E\{X\} = m$$

$$\mathrm{Var}\{X\} = m + \frac{m^2}{k} \tag{2}$$

so that $m$ is the mean of the distribution and $k$ can be interpreted as a measure of aggregation (Anscombe (1950) provides a nice discussion of these concepts). As $k$ increases, the level of aggregation decreases (for $k = \infty$, (2) shows that $E\{X\} = \mathrm{Var}\,\{X\}$, corresponding to the Poisson distribution). Values of $k$ for fish populations range from 0.1 up. Very often, $k$ is relatively constant if one samples the same species or the same physical scale (Taylor (1971)).

Observe from (1) that the probability of nonpositive sample (i.e., X=0, which will henceforth be called a negative sample) is

$$\Pr\{X=0\} = \left(\frac{k}{k+m}\right)^{k} \quad . \qquad (3)$$

Figure 1 shows Pr{X=0} as a function of $k$ and $m$. The impact of the figure is this: $m$ can be enormous (e.g., $10^7$) but if $k$ is small enough (e.g.,.1) there can still be a sizable chance (for $m = 10^7$, $k = .1$, it is .15) of a negative sample. The converse is true as well; when $k$ is small, a small fraction of the samples will drive the dynamics of the sample mean. That is, a few samples will yield enormous values of $X$. (For $k \approx .2$ to $.6$ the chance of an observation greater than twice the mean is about 15%; for an observation greater than 5 times the mean, the chance is 5% for k=.2, 4% for k=.4 and 2% for k=.6.) Table 1, for example, shows typical data on samples of anchovy larve. Clearly the few large samples will drive the dynamics of the sample mean.

Table 1

Data on 3.75 mm Anchovy Larvae in January, 1969

(from Zweifel and Smith, 1981).

| Station | Larve |
|---|---|
| 80.51 | 9 |
| .52 | 3 |
| .60 | 0 |
| 82.47 | 84 |
| 83.43 | 343 |
| .51 | 45 |
| .55 | 0 |
| 87.40 | 550 |
| .45 | 311 |
| .50 | 14 |
| .55 | 7 |
| 90.32 | 223 |
| .53 | 632 |
| 93.30 | 4 |
| .35 | 2 |
| .40 | 118 |
| .45 | 2 |
| 94.30 | 46 |
| 97.40 | 0 |

The level of aggregation of a species thus clearly effects design of a stock survey. In particular, the more highly aggregated a species, the more effort one would expect to put into survey for the same level of accuracy on the estimates. Leaman (1981) Hewitt, Smith and Brown (1976), Hewitt and Smith (1979), Hewitt (1976), and Hewitt and Smith (1982) discuss the problems of sample survey design. In particular, Hewitt and Smith (1982) study the effects of aggregation on the design of an acoustic survey and try to estimate the number of samples needed to achieve a given level of confidence.

This paper is concerned with the problem of sampling a population that has a NB distribution with known $k$ but unknown $m$. (The case of both $k$ and $m$ unknown is deferred to a later paper.) The kinds of questions that must be addressed are:

b) Given a discovery, say of $x_t$ in $n$ samples, what can one say about the distribution of $m$? What happens if one only pays attention to presence-absence of the sampled species?

2) What can one say about the confidence limits on $m$?

3) How many negative samples (i.e., zero counts) are needed before one can say with a given level of accuracy that m is less than a critical value $m_c$? The third question was the original motivation for this work; it is discussed in more detail on the next section.

In this paper, a Bayesian approach is advocated for two reasons. First, the Bayesian approach provides an objective method for incorporating information into the analysis. Second, the Bayesian approach is ideally suited for the use of "negative" information.

The kinds of sampling schemes discussed here are both fixed sample (i.e., take a fixed number of samples and then make inferences) and sequential sampling (Wald (1947), DeGroot (1970), Plant and Wilson (1985)) in which one consistently makes probability inferences. Sequential sampling schemes have the advantage of potentially requiring fewer samples. In addition, the methods developed here are simple and robust enough to be used in real time on the survey vessel if the vessel has a small microcomputer.

In the next section, the operational and biological motivations of this work (surveys for the Pacific Sardine) are discussed. The third and fourth sections concern presence-absence sampling in which the only data used are whether X=0 or X>0. The fifth section discusses a more traditional approach in which the value of X is actually used. In the sixth section, the problem of knowing when a habitat was exited is studied. The seventh section contains conclusions and directions for future work. There are three appendices. The first discusses the NB model and a new model (called Z/R) for aggregation. In the second appendix, the approximate noninformative prior for the mean of a negative binomial distribution is derived. In the third appendix, approximations for the integrals that arise in the paper are studied.

## OPERATIONAL AND BIOLOGICAL MOTIVATION: PACIFIC SARDINE MANAGEMENT

The work in this paper was motivated by the problem of designing egg surveys for the northern subpopulation of the Pacific Sardine (sardinops sagax, immortalized by John Steinbeck). This sardine population, which once may have been at a biomass greater than 11,000,000 metric tons (Smith 1978b) is now less than 20,000 short tons. Under the current management agreement, there will be no fishing on the sardine population until it exceeds 20,000 short tons, and egg surveys, which adapt the egg production method (Santander et al. (1982), Hewitt (1984), MacCall (1984), Wolf and Smith (1984)), will be used to determine whether the population exceeds the critical biomass level of 20,000 short tons or not. If there is sufficient evidence that the biomass exceeds the critical level, then a full biomass survey will be conducted and (possibly) some fishing pressure exerted.

The egg survey involves sampling areas of 0.05 $m^2$ at specified sites that are spaced 4nm by 10 nm apart. Figure 2 shows a proposed sardine egg sampling plan for a cruise in May, 1985 to be run by scientists at the Southwest Fisheries Center, La Jolla, California. Any eggs discovered at a sampling site are classified by age. Using historical data, Smith and Richardson (1977) estimated spawning biomass and the parameters $m$ and $k$ in the negative binomial distribution (1). They found $k$ relatively constant, between 0.1 and 0.2 as the spawning biomass varied by a factor of 40. Why should the eggs be so highly aggregated with a parameter that is virtually independent of spawning biomass? One explanation is the following (P. Smith, Southwest Fisheries Center, La Jolla, California). Regardless of spawning biomass, the eggs can get fertilized only if they (and the sperm) are sufficiently clumped at the time of fertilization.

That is, if the eggs are uniformly distributed, there is a considerable chance that most of them will not get fertilized. An observation consistent with this model is that eggs under 1 day old have $k \approx .2$, between 1 and 2 days old have $k \approx .4$ and greater than 3 days old have $k \approx .6$. This phenomenon is explained by assuming that after fertilization, the eggs begin to disperse, presumably due to random diffusion.

There is one more major biological difficulty -- one does not know if sampling site is actually a habitat for the sardine. (Smith 1978b, Fig. 98 shows historical spawning regions). Thus, if a negative sample is obtained, it can be because the site is not a habitat or because the site is a habitat and there were simply no eggs present. This leads one to a thorny operational problem. Superimposed upon the sampling sites on Figure 2 is a "habitat region." Hopefully, the habitat region is contained by the sampling plan, but the boundaries are unknown. Thus, as one traverses the sampling lines on Figure 2 from NE to SW, how does one know when the habitat has been exited?

## PRESENCE-ABSENCE SAMPLING WHEN ALL SITES ARE HABITATS

To begin, consider the problem of sampling when all sites are habitats, and the only data used are whether or not a site has eggs. (Although this problem is not really relevant for the sardine survey, it may be for other surveys and is thus included here for completeness.) It is assumed that corresponding to the critical biomass level is a critical value of $m$, $m_c$. That is, if the spawning biomass is less than the critical level, then $m < m_c$. The problem is to then determine the probability that $m < m_c$ as a function of the data.

Let $X_i$ denote the $i^{th}$ observation. When all sites are habitats, one has

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

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

Assume that $N$ samples are taken with $N_p$ of them positive and $N_n$ of them negative (and $N = N_p + N_n$). Let $\mathcal{L}(N_n,N_p|m)$ denote the likelihood of the sample, given a value of $m$, so that

$$\mathcal{L}\,(N_n,N_p|m) = \left(\frac{k}{k+m}\right)^{N_n k}\left[1-\left(\frac{k}{k+m}\right)^k\right]^{N_p} \quad . \qquad (5)$$

In order to use a Bayesian approach, one must specify a prior density $f_0(m)$. Two choices are the uniform prior (UP)

$$f_0(m) = 1,\ m \geq 0 \qquad (6)$$

and the approximate noninformative prior (NP)

$$f_0(m) = m^{-1/2}(k+m)^{-1/2},\ m \geq 0 \qquad (7)$$

(The NP (7) is derived in Appendix 2. A prior is noninformative if the data change only the location, but not the shape of the posterior density.) The UP given in (6) and NP given in (7) are improper in that they cannot be integrated on $[0,\infty]$. It will be seen that the posterior densities can be integrated under very general conditions. Figure 3 shows the prior densities UP and NP. Each has a certain attraction: UP because it essentially allows one to say "I don't know the value of $m$ and I initially give all values equal weight" and NP because of its property related to the data. Note that NP weights smaller values of $m$ more highly.

The posterior density on $m$ given $N_n$ and $N_p$, $f(m|N_n,N_p)$ is computed by Bayes Theorem

$$f(m|N_n,N_p) = \frac{f_0(m)\,\mathcal{L}(N_n,N_p|m)}{\int f_0(m)\,\mathcal{L}(N_n,N_p|m)dm} \qquad (8)$$

The posterior probability that $m \leq m_c$, $P(m \leq m_c)$, is then

$$P(m \leq m_c) = \int_0^{m_c} f(m|N_n,N_p)dm \qquad (9)$$

In order to compute (8) and (9), it is helpful to rewrite the likelihood (5) as follows

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

$$= k^{N_n k} \sum_{j=0}^{N_p} \binom{N_p}{j}(-1)^j k^{kj}(k+m)^{-k(N_n+j)} \qquad (10)$$

Let $P_{UP}(m \leq m_c)$ denote the posterior probability that $m \leq m_c$ when the uniform prior is used. Then

$$P_{UP}(m \leq m_c) = \frac{\sum_{j=0}^{N_p} \binom{N_p}{j}(-1)^j k^{kj}\int_0^{m_c}(k+m)^{-k(N_n+j)}dm}{\sum_{j=0}^{N_p} \binom{N_p}{j}(-1)^j k^{kj}\int_0^{\infty}(k+m)^{-k(N_p+j)}dm} \qquad (11)$$

In order to insure that the integrals in (11) converge, one should sample until $N_n > 1/k$. Then

$$P_{UP}(m \leq m_c) = \frac{\sum_{j=0}^{N_p} \binom{N_p}{j}(-1)^j \frac{k^{kj}}{k(N_n+j)-1}\left\{k^{-k(N_n+j)+1}-(k+m_c)^{-k(N_n+j)+1}\right\}}{\sum_{j=0}^{N_p} \binom{N_p}{j}(-1)^j \frac{k^{kj}}{k(N_n+j)-1}\, k^{-k(N_n+j)+1}}$$

The expression (12) is easily computed on a desktop microcomputer.

Let $P_{NP}(m \leq m_c)$ denote the posterior probability that $m \leq m_c$ when the noninformative prior is used. Then

$$P_{NP}(m \leq m_c) = \frac{\sum_{j=0}^{N_p} \binom{N_p}{j}(-1)^j k^{kj}\int_0^{m_c} m^{-1/2}(k+m)^{-k(N_n+j)-1/2}dm}{\sum_{j=0}^{N_p} \binom{N_p}{j}(-1)^j k^{kj}\int_0^{\infty} m^{-1/2}(k+m)^{-k(N_n+j)-1/2}dm} \qquad (13)$$

The integrals in (13) are most easily computed by setting

$$\left.\begin{aligned} m &= k\tan^2\theta \\ dm &= 2k\tan\theta d\theta/\cos^2\theta \\ k+m &= k/\cos^2\theta \end{aligned}\right\} \qquad (14)$$

and defining

$$\theta_c = \text{arc tan}\left(\sqrt{\frac{m_c}{k}}\right). \qquad (15)$$

When this is done, one finds that

$$P_{NP}(m \leq m_c) = \frac{\sum_{j=0}^{N_p}\binom{N_p}{j}(-1)^j k^{kj}\int_0^{\theta_c}(\cos\theta)^{2k(N_n+j)-1}d\theta}{\sum_{j=0}^{N_p}\binom{N_p}{j}(-1)^j k^{kj}\int_0^{\pi/2}(\cos\theta)^{2k(N_n+j)-1}d\theta} \qquad (16)$$

In order to insure convergence of the integrals in (16), one should sample until $N_n > 1/2k$. Although slightly more complicated than (12), (16) is also easily evaluated on a desktop microcomputer.

A discussion of sequential sampling plans and other probability statements is deferred until the next section when the generalization of this problem is treated.

## PRESENCE-ABSENCE SAMPLING WHEN SOME SITES ARE HABITATS

Now consider the case in which the $i^{th}$ site has a probability $P_i$ of being a habitat. Then, if a site is not a habitat $X_i = 0$ with probability 1. Consequently,

$$\left.\begin{aligned} \Pr\{X_i=0\} &= 1-P_i+P_i\left(\frac{k}{k+m}\right)^k \\ \Pr\{X_i>0\} &= P_i-P_i\left(\frac{k}{k+m}\right)^k \end{aligned}\right\} \qquad (17)$$

(Pennington (1983) and Pennington and Berrien (1984) discuss similar problems, but from a very different perspective.) Again, assume that there are $N_n$ negative samples and $N_p$ positive samples. Let $\eta$ denote those sites at which a negative sample was obtained and $\rho$ denote those sites at which a positive sample was obtained. The likelihood is then

$$\mathcal{L}(N_n,N_p|m) = \prod_{i\varepsilon\eta} \left\{1-P_i+P_i\left(\frac{k}{k+m}\right)^k\right\} \prod_{i\varepsilon\rho} \left\{P_i-P_i\left(\frac{k}{k+m}\right)^k\right\} \qquad (18)$$

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 (18) becomes

$$\mathcal{L}(N_n,N_p|m) = \left[1-P+P\left(\frac{k}{k+m}\right)^k\right]^{N_n}\left[P-P\left(\frac{k}{k+m}\right)^k\right]^{N_p} \qquad (19)$$

Consider first a simple probability statement that, after the $N_n$ negative and $N_p$ positive observations, $m \geq m_c$. It becomes convenient at this point to introduce a maximum allowed value of m, denoted by $m_m$. The value of

$m_m$, like the value of $m_c$, can be viewed as a (subjective) user input. When the uniform prior is used, one has

$$P_{UP}(m \geq m_c) = \frac{\int_{m_c}^{m_m} \left[1-P+P\left(\frac{k}{k+m}\right)^k\right]^{N_n} \left[P-P\left(\frac{k}{k+m}\right)^k\right]^{N_p} dm}{\int_{0}^{m_m} \left[1-P+P\left(\frac{k}{k+m}\right)^k\right]^{N_n} \left[P-P\left(\frac{k}{k+m}\right)^k\right]^{N_p} dm} \qquad (20)$$

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

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

The integral in (20) becomes

$$P_{UP}(m \geq m_c) = \frac{\int_{w_m}^{w_c} [1-P+Pw^k]^{N_p} [P-Pw^k]^{N_p} \frac{dw}{w^2}}{\int_{w_m}^{1} [1-P+Pw^k]^{N_p} [P-Pw^k]^{N_p} \frac{dw}{w^2}} \quad .$$

These integrals are easily computed on a desktop microcomputer. In addition, in Appendix 3 Gaussian approximations to the integrals are discussed.

When the noninformative prior and the transformation (14) are used, one obtains

$$P_{NP}(m \geq m_c) = \frac{\int_{\theta_c}^{\theta_m} [1-P+P(\cos\,\theta)^{2k}]^{N_n} [P-P(\cos\,\theta)^{2k}]^{N_p} \frac{d\theta}{\cos\,\theta^2}}{\int_0^{\theta_m} [1-P+P(\cos\,\theta)^{2k}]^{N_n} [P-P(\cos\,\theta)^{2k}]^{N_p} \frac{d\theta}{\cos\,\theta^2}} \qquad (21)$$

Figure 4 shows $\Pr\{m \geq m_c\}$ as a function of the number of positive samples using both priors. The NP is more "conservative" than the UP.

Figure 4 is a 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 5 is a sequential sampling diagram used to compute the probability that $m \leq m_c$ under the uniform prior. In this diagram one plots $N_n$ versus $N = N_p+N_n$. If an observation falls in the shaded region then one can conclude that $m \leq m_c$ with probability .95 (Fig. 5a) or .99 (Fig. 5b). If the current data point (N, $N_n$) 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 (18) is used. For example, when the UP is used one finds

$$P_{UP}(m \leq m_c) = \frac{\int_{w_c}^{1} \prod_{i \in n} \{1-P_i+P_iW^k\} \prod_{i \in p} (P_i-P_iW^k) \frac{dw}{w^2}}{\int_{w_m}^{w_c} \prod_{i \in n} \{1-P_i+P_iW^k\} \prod_{i \in p} (P_i-P_iW^k) \frac{dw}{w^2}} \qquad (22)$$

The only difficulty is that one cannot develop charts similar to Figures 4 and 5. On the other hand, (22) 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 (22) shows that $P_{UP}(m \leq m_c) = .52$. Suppose that the next five data points are (1,0), (.95, 1), (.95,0), (.95,0) and (.9,0). Using the 15 data points gives $P_{UP}(m \leq m_c) = .82$. If the next five data points are (1,0), (1,0), (.95,0), (.95,0), and (.9,1) then $P_{UP}(m \leq m_c) = .86$. If the next five data points are (1,0), (1,0), (.95,0), (.95,0) and (.9,0) then $P_{UP}(m \leq m_c) = .96$ 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. For example, in the sardine cruise depicted in Figure 2, only 15 of the 300 sites may have eggs (pers. comm. P. Smith, SWFC, La Jolla, CA). Second, a large amount of negative information is needed to insure that $m \leq m_c$ with a high confidence level. One must remember, however, that with the UP, the initial probability that $m \leq m_c$ is ${}^{m}c/m_m$. So, for example, for the values presented here the prior probability that $m \leq m_c$ is $1.14 \times 10^{-3}$. In addition, since not every site is a habitat the effects of negative information on the updated distribution are mitigated (i.e., as $P_i \to 0$ the data have decreasing effects on the Bayesian update).

## BAYESIAN UPDATING OF THE MEAN

Now consider the situation in which the actual values of $X_i$ are used to update the distribution of the mean. Observe from (1) that if every site is a habitat, then

$$\Pr\{X=x\} \propto \frac{m^x}{(k+m)^{k+x}} \qquad (23)$$

where the proportionality constant contains terms independent of m. Consequently, if N samples are taken, with $X_i$ having the value $x_i$ and $x_T = \sum_{i=1}^{N} x_i$, the likelihood of the data is

$$\mathcal{L} = m^{x_T}(k+m)^{-Nk-x_T} \qquad (24)$$

If the prior density $f_0(m)$ is assumed to be of the form

$$f_0(m) \propto m^{\alpha}(k+m)^{\beta} \qquad (25)$$

then the posterior density is

$$f(m|x) \propto m^{\alpha+x_T}(k+m)^{-Nk-x_T+\beta}. \qquad (26)$$

The uniform prior corresponds to $\alpha=\beta=0$ and the noninformative prior corresponds to $\alpha=\beta=-1/2$. The posterior density (26) can be integrated using one of the two substitutions discussed in the previous section.

For purposes of sequential sampling, observe that (24) - (26) can be summarized as follows. At any point in the sampling scheme, the density of m is proportional to

$$m^{\alpha}(k+m)^{\beta} \qquad (27)$$

If the next sample has the value x, the updated density has the same form as (27) with updated parameters $\alpha'$, $\beta'$ given by

$$\begin{aligned} \alpha' &= \alpha+x \\ \beta' &= \beta-k-x \end{aligned} \tag{28}$$

Next consider the case in which the $i^{th}$ site is a habitat with probability $P_i$. Then

$$\left.\begin{aligned} &Pr\{X_i=0\} = 1-P_i+P_i(\frac{k}{k+m})^k \\ \\ &Pr\{X_i= x_i>0\} \propto P_i m^{x_i}(k+m)^{-k-x_i} \end{aligned}\right\} \tag{29}$$

As before, let $\eta$ denote those $i$ for which $X_i= 0$ and $\rho$ those $i$ for which $X_i > 0$. The likelihood of the data is then

$$\mathcal{L} \propto \prod_{i\epsilon\eta} \{1-P_i+P_i(\frac{k}{k+m})^k] \prod_{i\epsilon\rho} [P_i m^{x_i}(k+m)^{-k-x_i}] \tag{30}$$

If $\rho$ contains p data points, $x_p= \sum_{i\epsilon\rho} x_i$, and the prior density (25) is used, the posterior density is

$$f(m|X) \propto m^{\alpha}(k+m)^{\beta}\{\prod_{i\epsilon\eta} [1-P_i+P_i(\frac{k}{k+m})^k\}\{\prod_{i\epsilon\rho} P_i m^{x_i}(k+m)^{-k-x_i}\}$$

$$=\{\prod_{i\epsilon\eta} [1-P_i+P_i(\frac{k}{k+m})^k\}\{\prod_{i\epsilon\rho} P_i\} m^{\alpha+x_p}(k+m)^{-pk-x_p+\beta}. \tag{31}$$

Equations (26) and (31) provide a complete Bayesian method for updating the mean. The only difficulty involved may be one in which computational problems occur, due to exponentiation to very high powers (caused by large values of $x_T$ or $x_p$). This problem aside, one can redo all of the kinds of sampling plans that were done in the previous section.

## EXIT FROM A HABITAT

Next, consider the problem of estimating when the habitat has been exited. That is, suppose a string of negative samples has been obtained. What is the probability that one is sampling a region in which the sites are simply not habitats? In order to formulate this problem, a habitat profile is needed. That is, define a "site" variable $s$ and let $p(s)$ denote the probability that a site at point $s$ is a habitat. Panels a) -c) of Figure 6 show three possible habitat profiles. These are respectively

$$p(s) = \begin{cases} 1 - s/s_0 & s < s_0 \\ 0 & s \geq s_0 \end{cases} \qquad (32a)$$

$$p(s) = \begin{cases} 1 & s < s_0 \\ 0 & s \geq s_0 \end{cases} \qquad (32b)$$

$$p(s) = \begin{cases} 1-\alpha s & s < s_0 \\ 0 & s \geq s_0 \end{cases} \qquad (32c)$$

The first two profiles are single parameter models in which the parameter $s_0$ must be estimated; the last profile is a two parameter model in which $s_0$ and $\alpha$ must be estimated. For the sardine eggs, at least, no gradations of habitat appear to be observed (P. Smith, personal communication, SWFC, La Jolla, California) so that profile (32b) is the one of choice.

The set up for Bayesian updating is shown in Figure 6d. Let $s$ denote the current position of the vessel, $s_1$ the first negative sample in the current string and $\Delta$ the distance between sites (so that the last positive sample was at $s_1 - \Delta$). One wishes to compute the posterior probability that $s_0 \leqq s$, given the string of negative samples. For the habitat profile (32b), one has

$$\Pr[X_i = 0\} = \begin{cases} (\frac{k}{k+m})^k & \text{if} \quad i < s_0 \\ 1 & \text{if} \quad i \geq s_0 \end{cases} \qquad (33)$$

Consequently, the likelihood of the data, conditioned on $s_0 = s$, is

$$\mathcal{L}\ \{\text{data}|s_0=s\} = (\frac{k}{k+m})^{k[\frac{s-s_1}{\Delta}]} \qquad (34)$$

If $f_0(s)$ is the prior probability density for $s_0$, the posterior density is

$$\frac{f_0(s)(\frac{k}{k+m})^{ks/\Delta}}{[\int f_0(s)(\frac{k}{k+m})^{ks/\Delta} ds]}, \ s \geq s_1 \qquad (35)$$

For example, a relatively robust and versatile choice of prior densities is the family of gamma densities

$$f_0(s) = \frac{\alpha^\nu}{\Gamma(\nu)}\ e^{-\alpha s}\ s^{\nu-1} \qquad (36)$$

with two parameters $\nu$ and $\alpha$. For $\nu = 1$, (36) is an exponential density, which is rewritten as

$$f_0(s) = \frac{1}{\bar{s}}\ e^{-s/\bar{s}} \qquad (37)$$

where $\bar{s}$ is the prior estimate of the mean of $s_0$. A little calculation shows that

$$\Pr\{s_0 < s|\text{data}\} = 1 - e^{-(s-s_1)/\lambda} \qquad s \geq s_1 \qquad (38)$$

where

$$\frac{1}{\lambda} = \frac{1}{\bar{s}} - \frac{k}{\Delta} \log\ (\frac{k}{k+m}) \qquad (39)$$

For the general gamma density (36), one finds after some calculation that

$$\Pr\{s_0 < s|\text{data}\} = \frac{\gamma(\nu,\ (\alpha+\tilde{k})s)-\gamma(\nu,(\alpha+\tilde{k})^{s}1)}{\Gamma(\nu) - \gamma(\nu,(\alpha+\tilde{k})s_1)} \qquad s \geq s_1 \tag{40}$$

where $\tilde{k} = (-k/\Delta) \log (k/k+m)$ and $\gamma(\nu,x)$ is the incomplete gamma function defined by

$$\gamma(\nu,x) = \int_0^x e^{-t}t^{\nu-1}dt = \frac{\Gamma(\nu)}{x^{-\nu}}\, e^{-x} \sum_{n=0}^{\infty} \frac{x^n}{\Gamma(\nu+n+1)} \tag{41}$$

Observe that in order to use (38) or (40) one must specify a value of m. (The problem of jointly estimating m and whether one is inside or outside of a habitat is a more difficult problem, deferred to a later paper). Also, observe that if $\bar{s}$ and $\nu$ are given, then

$$\frac{\nu}{\alpha} = \bar{s} \tag{42}$$

so that $\alpha = \nu/\bar{s}$. Table 2 shows sample output using (38) and (40). For the second set of data, when $\nu=.25$, the posterior probability that $s < s_0$ exceeds .99 only for $s > 104$.

## Table 2

## Exiting the Habitat

Data $k=.2$ $m=1$ $\bar{s}=20$ $s_1=12$ $\Delta=4$

$\Pr\{s_0 < s\}$ for $\nu$ =

| s | 2(CV=.71) | 1 (CV=1) | 1/2 (CV=1.414) | 1/4 (CV=2) |
|---|---|---|---|---|
| 12 | 0 | 0 | 0 | 0 |
| 16 | .42 | .43 | .43 | .43 |
| 20 | .68 | .67 | .67 | .67 |
| 24 | .83 | .81 | .80 | .80 |
| 28 | .91 | .89 | .88 | .88 |
| 32 | .95 | .94 | .93 | .93 |
| 36 | .97 | .96 | .96 | .95 |
| 40 | .99 | .98 | .97 | .97 |
| 44 | .99 | .99 | .98 | .98 |
| 48 | .99 | .99 | .99 | .99 |
| 52 | .99 | .99 | .99 | .99 |

Table 2 (continued)

Data $k=.2$ $m=1$ $\bar{s}=20$ $s_1=12$ $\Delta=4$

$\Pr\{s_0 < s\}$ for $\nu$ =

| s | 1 | .25 |
|---|---|---|
| 12 | 0 | 0 |
| 16 | .25 | .22 |
| 20 | .43 | .38 |
| 24 | .57 | .50 |
| 28 | .68 | .59 |
| 32 | .75 | .66 |
| 36 | .81 | .72 |
| 40 | .86 | .77 |
| 44 | .89 | .81 |
| 48 | .92 | .84 |
| 52 | .94 | .86 |
| 56 | .95 | .89 |
| 60 | .97 | .90 |
| 64 | .97 | .92 |
| 68 | .98 | .93 |
| 72 | .99 | .94 |
| 76 | .99 | .95 |
| 80 | .99 | .96 |
| 84 | .99 | .96 |

## CONCLUSIONS AND DISCUSSION

In this paper, a number of different Bayesian methods for sampling highly aggregated popluations are introduced. There remain issues and questions associated with these models that are worth discussing. These include the following.

1. Which Technique Should Be Used? The three techniques introduced here are i) presence-absence sampling when all sites are habitats, ii) presence-absence sampling when some sites are habitats and iii) Bayesian updating of the mean. From the standpoint of both operational reality and computational feasibility, it appears that technique ii) is the one of choice. Operational realism is included by allowing sites to have a non zero probability of not being a habitat. By doing only presence-absence sampling one avoids the kinds of computational difficulties that were described in the fifth section. Finally, the software for technique ii) is easily developed and robust. The ultimate decision about the technique of choice must depend, to some extent, upon operational testing.

2. Many Age Classes of Eggs. In the actual sardine survey, three age classes of eggs (<1 day, 1-2 day, 2-3 days old) are sampled, each with a different aggregation parameter (k=.2, .4, .6 respectively). Thus, the data are more complicated, consisting of presence-absence of the three age classes or the actual counts of the three age classes. The question of how to use these data is complex. One could assume, for example, that the three 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) page 568) where the correlation level is a user inputted variable.

For the sardines, at least, biologists believe that the three age classes can indeed be treated as independent variables (P. Smith, Southwest Fisheries Center, La Jolla, CA). The reason for this belief is that the sardines are highly mobile and eggs of different ages are spawned by different schools. There is considerable acoustic evidence that different schools of sardines move independently.

If $m_j$ and $k_j$ now denote the values of m and k for the $j^{th}$ egg class, the independence assumption implies that the probability of no eggs at the $i^{th}$ sampling site is given by

$$\Pr\{X_i=0\} = \prod_j \{1-P_i+P_i(\frac{k_j}{k_j+m_j})^{k_j}\} \qquad (43)$$

All of the results presented in the previous sections can easily be modified using (43).

3. Joint Estimation of $s_0$ and m. One problem not discussed in any of the previous sections was the joint estimation of the extent of the habitat and the value of m. In principle, this joint estimation problem can be tackled using the same kinds of methods found in sections 3-6. It is likely, however, that the details will be more complex. For this reason, discussion of the joint estimation problem is deferred.

4. 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

$$\Pr\{X_i = x\} = \frac{\Gamma(k+x)}{x!\,\Gamma(k)} \left(\frac{k}{mW_i+k}\right)^k \left(\frac{mW_i}{mW_i+k}\right)^x \qquad (44)$$

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.

5. 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 easily 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.

6. 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.

# ACKNOWLEDGEMENTS

This work was partially supported by NSF Grant MCS-81-21659 and by the Agricultural Experiment Station of the University of California.

I thank Richard E. Plant for initial discussions about the non-informative prior for the negative binomial distribution and Paul E. Smith for numerous discussions about the egg sampling problem.

## REFERENCES

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

Bender, C.M. and S.A. Orszag. 1978. Advanced Mathematical Methods for Scientists and Engineers. McGraw-Hill, New York. 593 pg.

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

Bleistein, N. and R.A. Handelsman. 1975. Asymptotic Expansions of Integrals. Holt, Rinehart and Winston, New York. 425 pgs.

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

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

DeGroot, M. 1970. Optimal Statistical Decisions. Mc-Graw Hill, New York. 489 pgs.

Feller, W. 1968. An Introduction to Probability Theory and Its Applications, Vol. 1. John Wiley, New York. 509 pp.

Gerard, G. and P. Berthet. 1971. Sampling strategy in censusing patch populations pages 59-68 in Statistical Ecology, Volume 1 (G.P. patil, E.C. Pielou and W.E. Waters, eds) Pennsylvania State University Press, University Park, PA.

Gunderson, D.R., Thomas, G.L., Cullenberg, P., Eggers, D.M. 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 18:149-154.

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

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

Hewitt, R. and P. Smith. 1979. Seasonal distributions of epipelagic fish schools and fish biomass over portions of the California Current region. CalCOFI Rep. 20:102-110.

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 74:281-300.

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

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

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

McCall, A.D. 1984. Population Models of Habitat Selection, with Application to the Northern Anchovy. Administrative Report LJ-84-01, Southwest Fisheries Center, National Marine Fisheries Service, La Jolla, California 92038.

Mangel, M. 1984. Decision and Control in Uncertain Resource Systems. Academic Press, New York. 255 pgs.

Mangel, M. and J.H. Beder. 1985. Search and stock depletion: Theory and Applications. Can. J. Fish. Aq. Sci. 42:150-165

Mangel, M., Plant, R. E. and J.R. Carey. 1984. Rapid delimiting of pest infestations: A case study of the Mediterranean Fruit Fly, J. appl. Ecol. 21:563-579.

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

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

Pennington, M. and P. Berrien. 1984. Measuring the precision of estimates of total egg production based on plankton surveys. J. Plankton Res. 6(5):

Pielou, E.C. 1977. Mathematical Ecology. Wiley, New York. 385 pgs.

Plant, R.E. and T. Wilson. 1985. A Bayesian method for sequential sampling and forecasting in agricultural pest management. Biometrics, forthcoming.

Santander, H., Smith, P.E., and J. Alheit. 1982. Determination of sampling effort required for estimating egg production of anchovetta, Engrualis ringens, off Peru. Boletin Inst. del mar del Peru. 7:5-18.

Smith, P.E. 1978a. Precision of sonar mapping for pelagic fish assessment in the California Current. J. Cons. int. Explor. Mer. 38:3-40.

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

Smith, P. and S.L. Richardson. 1977. Standard Techniques for Pelagic Fish Egg and Larva Surveys. FAO Fisheries Technical Paper No. 175., Food and Agriculture Organization of the United Nations, Rome. 100 pgs.

Taylor, C.C. 1953. Nature of Variability in Trawl Catches. Fishery Bulletin 83, U.S. Department of the Interior. Volume 54:145-166.

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

Wald, A. 1947. Sequential analysis. Dover, New York. 212 pgs.

Wolf, P. and P. Smith. 1984. An inverse egg production method for objectively determining whether the northern subpopulation of Pacific Sardine is greater than or less than 20,000 tons spawning biomass on an annual basis. CalCOFI Abstract, October 1984.

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 178:248:259.

## APPENDIX 1. The NB and Z/R Models

In this appendix an operational motivation of the NB model is presented. It is used to motivate a new model, the Z/R (Zero/Random) model, for aggregation.

Suppose that at any site counts are random; thus they are given by the Poisson with parameter $\lambda$

$$\Pr\{x|\lambda\} = \frac{e^{-\lambda}\lambda^{x}}{x!} \qquad \text{(A-1)}$$

Assume next that $\lambda$ varies globally, so that in any given site the actual value of $\lambda$ is unknown. For a distribution on $\lambda$, pick the gamma density with parameters $k$ and $w$:

$$f(\lambda) = \frac{e^{-w\lambda}\lambda^{k-1}w^{k}}{\Gamma(k)} \qquad \text{(A-2)}$$

The unconditional distribution of counts is then

$$\Pr\{x\} = \int_0^\infty \frac{e^{-\lambda}\lambda^{x}}{x!}\;\frac{e^{-w\lambda}\lambda^{k-1}}{\Gamma(k)}\;w^{k}\,d\lambda$$

$$\text{(A-3)}$$

$$= \frac{\Gamma(k+x)}{x!\,\Gamma(k)}\left(\frac{w}{w+1}\right)^{k}\left(\frac{1}{w+1}\right)^{x}$$

Next, define $m$ by $m = k/w$. Then (A-3) can be rewritten as

$$\Pr\{x\} = \frac{\Gamma(k+x)}{x!\,\Gamma(k)}\left(\frac{k}{k+m}\right)^{k}\left(\frac{m}{m+k}\right)^{x} \qquad \text{(A-4)}$$

Equation (A-4) is the NB distribution (1).

This kind of calculation motivates the Z/R model. That is, assume that a given site has a probability $p$ of containing a positive number of eggs. If it contains eggs, assume that the number of eggs at the site is randomly distributed (i.e. Poisson). Thus

$$\Pr\{X_i = 0\} = 1-p$$

$$\Pr\{X_i = x\} = \frac{p}{e^{\lambda}-1}\ \frac{\lambda^x}{x!} \qquad ,x = 1,2... \qquad \text{(A-5)}$$

The $e^{\lambda}-1$ in the denominator arises because $\sum_{k=1}^{\infty} \lambda/k! = e^{\lambda}-1$. A straightforward calculation shows that

$$E\{X_i\} = \frac{p\lambda e^{\lambda}}{e^{\lambda}-1}$$

$$\text{(A-6)}$$

$$E\{X_i^2\} = \frac{pe^{\lambda}(\lambda+\lambda^2)}{e^{\lambda}-1}$$

Although an analytic form for the variance $Var\{X_i\}$ is not especially tractable, it is easily computed. Figure A-1 shows a plot of $Var\{X_i\}/E\{X_i\}$ as a function of the mean $E\{X_i\}$. Clearly, very large variance-mean ratios can be observed, as in the case of the NB distribution.

For this reason, it is worth considering inference for the Z/R model. Suppose that of N samples, n are negative and N-n are non-zero, with $X_i$ denoting the value of the $i^{th}$ sample. The likelihood, $\hat{\mathcal{L}}$, of such a data set is

$$\hat{\mathcal{L}} = (1-p)^n\ p^{N-n}\ \frac{1}{(e^{\lambda}-1)}\ N\text{-}n\ \Pi\ \frac{\lambda^{x_i}}{x_i!} \qquad \text{(A-7)}$$

so that the log-likelihood, $\mathcal{L}$, is

$$\mathcal{L} = n\log(1-p) + (N-n)p-(N-n)\log(e^{\lambda}-1)$$

$$+ \sum x_i \log \lambda - x_i!. \qquad \text{(A-8)}$$

Differentiating L with respect to p gives

$$\frac{\partial \mathcal{L}}{\partial p} = \frac{-n}{1-p} + \frac{N-n}{p}$$

$$\text{(A-9)}$$

$$\frac{\partial^2 \mathcal{L}}{\partial p^2} = \frac{-n}{(1-p)^2} - \frac{N-n}{p^2}$$

Thus the MLE $\hat{p}$ is given by

$$\hat{p} = \frac{N-n}{N} \qquad \text{(A-10)}$$

and

$$\frac{\partial^2 \mathcal{L}}{\partial p^2} \Big|_{\hat{p}} = \frac{-N^3}{n(N-n)} \qquad \text{(A-11)}$$

Differentiating (A-8) with respect to $\lambda$ gives

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \frac{(N-n)}{e^{\lambda}-1} e^{\lambda} + \sum \frac{x_i}{\lambda}$$

$$\text{(A-12)}$$

$$\frac{\partial^2 \mathcal{L}}{\partial \lambda^2} = \frac{-x_t}{\lambda^2} + \frac{(N-n)e^{-\lambda}}{(1-e^{-\lambda})^2}$$

where $x_t = \sum\limits_i x_i$. Thus, the MLE for $\lambda$ satisfies

$$\left(\frac{1}{N-n}\right) x_t = \frac{\lambda e^{\lambda}}{e^{\lambda}-1} = \frac{\lambda}{1-e^{-\lambda}} \qquad \text{(A-13)}$$

Equation (A-13) is easily solved numerically. As an example, consider the egg data shown in Table A.1, corresponding to anchovy per $10m^2$. There were 208 zeroes.

The sample mean is 206, the sample standard deviation in the 419 samples is 1022 and $x_t$ = 86120. Solution of equations (A-9) to (A-13) shows that

$$\hat{p} = .504$$

$$\hat{\lambda} = 408 \tag{A-14}$$

Manipulating such large values of $\lambda$ in (A-5) is difficult. Hence, it helps to rewrite (A-5) as

$$\begin{aligned} \Pr\{X_i=x\} &= \frac{p}{1-e^{-\lambda}} \frac{\lambda^x e^{-\lambda}}{x!} \\ &= \frac{p}{1-e^{-\lambda}} \frac{1}{x!} \exp\{x\log\lambda-\lambda\} \end{aligned} \tag{A-15}$$

Further investigation of the Z/R model seems warranted since it has such a nice conceptual base.

## TABLE A-1

Sample Size: 419

Number of Zeroes: 208

Positive Data

| $X_1$ | Number of Occurrences | $X_i$ | Number of Occurrences |
|---|---|---|---|
| 2 | 6 | 58 | 1 |
| 3 | 30 | 60 | 1 |
| 4 | 8 | 62 | 1 |
| 5 | 4 | 63 | 1 |
| 6 | 13 | 66 | 1 |
| 7 | 3 | 71 | 1 |
| 8 | 7 | 72 | 1 |
| 9 | 6 | 73 | 1 |
| 10 | 1 | 86 | 1 |
| 11 | 5 | 90 | 1 |
| 12 | 5 | 93 | 1 |
| 13 | 2 | 95 | 1 |
| 14 | 2 | 97 | 1 |
| 15 | 4 | 104 | 2 |
| 16 | 3 | 111 | 1 |
| 17 | 1 | 119 | 2 |
| 18 | 2 | 120 | 1 |
| 19 | 1 | 123 | 1 |
| 20 | 1 | 125 | 1 |
| 22 | 3 | 156 | 1 |
| 23 | 3 | 157 | 1 |
| 24 | 1 | 164 | 1 |
| 25 | 1 | 167 | 1 |
| 26 | 1 | 175 | 1 |
| 27 | 1 | 217 | 1 |
| 28 | 1 | 225 | 1 |
| 30 | 1 | 227 | 1 |
| 31 | 1 | 230 | 1 |
| 33 | 1 | 237 | 1 |
| 34 | 1 | 243 | 1 |
| 37 | 1 | 275 | 1 |
| 40 | 1 | 314 | 1 |
| 42 | 2 | 315 | 2 |
| 43 | 1 | 381 | 1 |
| 47 | 2 | 389 | 1 |
| 48 | 1 | 411 | 1 |
| 51 | 1 | 448 | 2 |
| 56 | 1 | 456 | 1 |
|  |  | 487 | 1 |
|  |  | 524 | 1 |
|  |  | 540 | 2 |
|  |  | 562 | 1 |

| $X_i$ | Number of Occurrences |
|---|---|
| 632 | 1 |
| 634 | 1 |
| 745 | 1 |
| 747 | 1 |
| 791 | 1 |
| 964 | 1 |
| 1025 | 1 |
| 1058 | 1 |
| 1061 | 1 |
| 1155 | 1 |
| 1315 | 1 |
| 1388 | 1 |
| 1484 | 1 |
| 1535 | 1 |
| 1613 | 1 |
| 1838 | 1 |
| 1851 | 1 |
| 1856 | 1 |
| 2506 | 1 |
| 2874 | 1 |
| 4442 | 1 |
| 5819 | 1 |
| 6619 | 1 |
| 8390 | 1 |
| 9488 | 1 |
| 12232 | 1 |

## APPENDIX 2. 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 (1) s the likelihood of m given x, the log-likelihood is

$$L(m|x) = -k\log(k+m) + x[\log m - \log(m+k)] + \ell(x,k) \qquad \text{(A-16)}$$

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

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

$$\text{(A-17)}$$

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

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

$$J(\hat{m}) = -\frac{\partial^2 L}{\partial m^2}\Big|_{\hat{m}}$$

$$= \frac{\hat{m}}{\hat{m}^2} - \frac{\hat{m}+k}{(k+\hat{m})^2} = \frac{k}{\hat{m}(k+\hat{m})} \qquad \text{(A-18)}$$

The approximate non-informative prior is then

$$f_0(m) \propto J(m)^{1/2} = m^{-1/2}(k+m)^{-1/2} \qquad \text{(A-19)}$$

## APPENDIX 3: Gaussian Approximations for the Posterior Distributions

In this appendix, various methods for approximating some of the integrals that arise in the body of the paper are discussed. The typical integral (analogous to (26)) is

$$P_{UP}(m \leq m_c) = \frac{\int_{w_c}^{1}[1-P+Pw^k]^{N_n}[P-Pw^k]^{N_p}\frac{dw}{w^2}}{\int_{w_m}^{1}[1-P+Pw^k]^{N_n}[P-Pw^k]^{N_p}\frac{dw}{w^2}} \qquad \text{(A-20)}$$

Since $N_p = N - N_n$, the numerator (for example) in (A-20) can be rewritten as

$$I = \int_{w_c}^{1}\frac{1}{w^2}\exp\left\{(N-N_n)\log(P-Pw^k)+N_n\log(1-P+Pw^k)\right\}dw \qquad \text{(A-21)}$$

Now set $N_n = fN$, where $0 \leq f \geq 1$, to obtain

$$I = \int_{w}^{1}\frac{1}{w^2}\exp\left\{N\left[\log(P-Pw^k) + f\log(1-P+Pw^k)\right.\right.$$

$$\left.\left.-f\log(P-Pw^k)\right]\right\}dw$$

$$= \int_{w_c}^{1}\frac{1}{w^2}\exp\left\{N\left[\log(P-Pw^k) + f\log\left(\frac{1-P+Pw^k}{P-Pw^k}\right)\right]\right\}dw \qquad \text{(A-22)}$$

If $N$ is large (as it will be in many applications) then (A-22) is of the form of a Laplace integral and can be analyzed accordingly (Bleistein and Handelsman (1975), Bender and Orszag (1978)).

To do this, define

$$Q(w,f) = \log(P-Pw^k) + f\log\left\{\frac{1-P+Pw^k}{P-Pw^k}\right\} \qquad \text{(A-23)}$$

so that

$$I = \int_{w_c}^{1} \frac{1}{w^2} \exp\left[N\, Q(w,f)\right] \, dw \qquad \text{(A-24)}$$

The main contribution to the integral I will come from the vicinity of the maximum of Q(w,f). If f is large enough, the equation

$$\frac{\partial Q}{\partial w} = 0 \qquad \text{(A-25)}$$

has a solution $w_0$. It is easy to show that $Q(w_0,f)$ corresponds to a maximum of Q(w,f). Thus, let

$$Q_s = \left|\frac{\partial^2 Q}{\partial w^2}\right|_{w_0}. \qquad \text{(A-26)}$$

Now consider the generalization of (A-24) to

$$I = \int_{w_e}^{1} \frac{1}{w^2} \exp\left[NQ(w,f)\right]dw \qquad \text{(A-27)}$$

where $w_e$ denotes an end point, either $w_e = w_c$ or $w_e = w_m$. Two cases need to be considered.

Case 1 $w_e < w_0 < 1$ and $Q(w_0,f) > Q(w_e,f)$

(Note that $Q(w,f) \rightarrow -\infty$ as $w \rightarrow 1$). This case corresponds to an internal maximum of Q(w,f). One proceeds as follows.

$$I = \int_{w_e}^{1} \frac{1}{w^2} \exp\{NQ(w;f)\}\, dw$$

$$\approx \frac{\exp\{NQ(w_0;f)\}}{w_0^2} \int_{w_e}^{1} \exp\{-N\frac{Q_s}{2}(w-w_0)^2\}\, dw \qquad \text{(A-28)}$$

Now set

$$y^2 = NQ_s(w-w_0)^2$$

$$y = \sqrt{NQ}_s(w-w_0) \tag{A-29}$$

$$dy = \sqrt{NQ}_s\ dw$$

Using (A-29) gives

$$I \approx \frac{\sqrt{2\pi}\ \exp\{NQ(w_0;f)\}}{\sqrt{NQ}_s\ w_0^2} \int_{\sqrt{NQ}_s(w_e-w_0)}^{\sqrt{NQ}_s(1-w_0)} e^{-y^2/2} \frac{dy}{\sqrt{2\pi}} \tag{A-30}$$

$$= \sqrt{\frac{2\pi}{NQ_s}}\ \frac{\exp\{NQ(w_0;f)}{w_0{}^2} \left[\Phi\ (\sqrt{NQ}_s(1-w_0)) - \Phi(\sqrt{NQ}_s(w_e-w_0))\right] \tag{A-31}$$

where

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-y^2/2}\ dy \tag{A-32}$$

is the cumulative distribution function for the Gaussian density. Using (A-31), the Gaussian approximation to (A-20) is

$$P_{UP}(m \leq m_c) \approx \frac{\Phi(\sqrt{NQ}_s(1-w_0)) - \Phi(\sqrt{NQ}_s(w_c-w_0))}{\Phi(\sqrt{NQ}_s(1-w_0)) - \Phi(\sqrt{NQ}_s(w_m-w_0))} \tag{A-33}$$

Case 2 $w_0 \varepsilon\ [w_e,1]$ or $Q(w_e) > Q(w_0)$

In this case, the main contribution to the integral (A-27) comes from the end point $w_e$. One proceeds as follows.

$$I = \int_{w_e}^{1} \frac{1}{w^2} \exp\{NQ(w;f)\}\, dw \qquad \text{(A-34)}$$

$$\approx \frac{\exp\{NQ(w_e,f)\}}{w_e^2} \int_{w_e}^{1} \exp\,\{NQ_w(w_e;f)(w-w_e)\}\, dw \qquad \text{(A-35)}$$

Now set

$$q = -\left.\frac{\partial Q}{\partial w}\right|_{w_e}$$

to obtain

$$I \approx \frac{\exp\{NQ(w_e;f)\}}{w_e^2} \int_{w_e}^{1} \exp\{-Nq(w-w_e)\}dw \qquad \text{(A-36)}$$

Setting $y = Nq(w-w_e)$, $dy = Nqdw$ gives

$$I \approx \frac{\exp\{NQ(w_e;f)\}}{Nq\ {w_e}^2} \int_{0}^{Nq(1-w_e)} e^{-y}dy$$

$$= \frac{\exp\{NQ(w_e;f)\}}{Nq{w_e}^2}\,(1-e^{-Nq(1-w_e)}) \qquad \text{(A-37)}$$

With this approximation, one has

$$P_{UP}(m \le m_c) \approx \frac{e^{NQ(w_c;f)}\,(1-e^{-Nq(1-w_c)})}{e^{NQ(w_m;f)}\,(1-e^{-Nq(1-w_m)})} \qquad \text{(A-38)}$$

The advantage of (A-34) or (A-38) is clear: They no longer involve numerical integration. The most that one has to do is numerically solve the nonlinear equation (A-25) and evaluate the cumulative Gaussian distribution.

CAPTIONS FOR FIGURES

Figure 1. Likelihood of a zero observation in a NB distribution with parameters m and k.

Figure 2. The sampling sites for the 1985 sardine egg survey proposed by NMFS Scientists (taken from Wolf and Smith (1984)).

Figure 3. Uniform (UP) and noninformative (NP) priors $f_0(m)$.

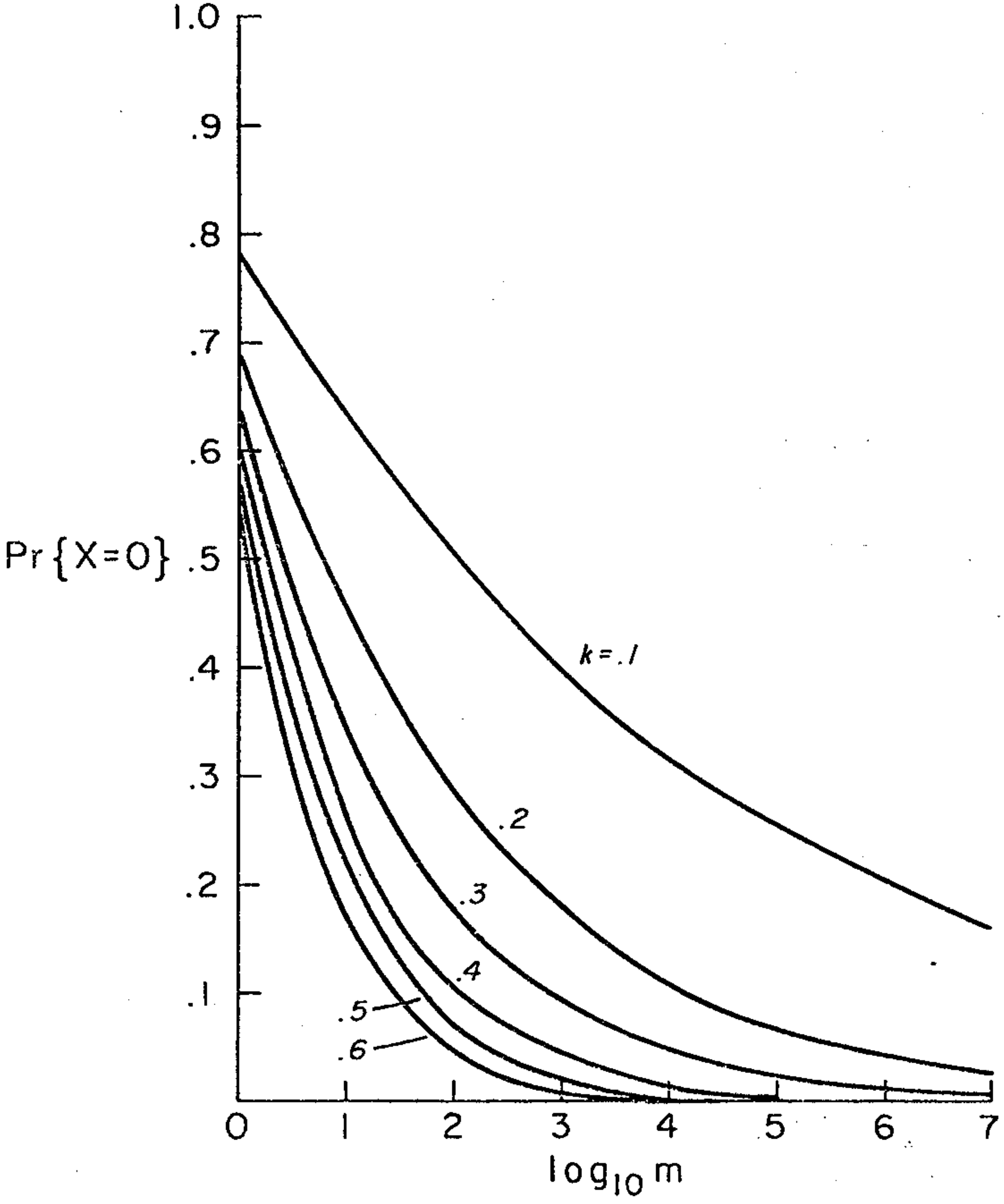
Figure 4. Probability that m exceeds $m_c$=1.14 as a function of the number of positive samples ($N_p$) in a total of N=100 samples. Other parameters: $m_m$=1000, k=.2, p=.8.

Figure 5. Sequential sampling charts in which the number of negative samples ($N_n$) is plotted against the total number of samples. If the data fall in the shaded region, one can conclude with 95% confidence (Figure 5a) or 99% confidence (Figure 5b) 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.

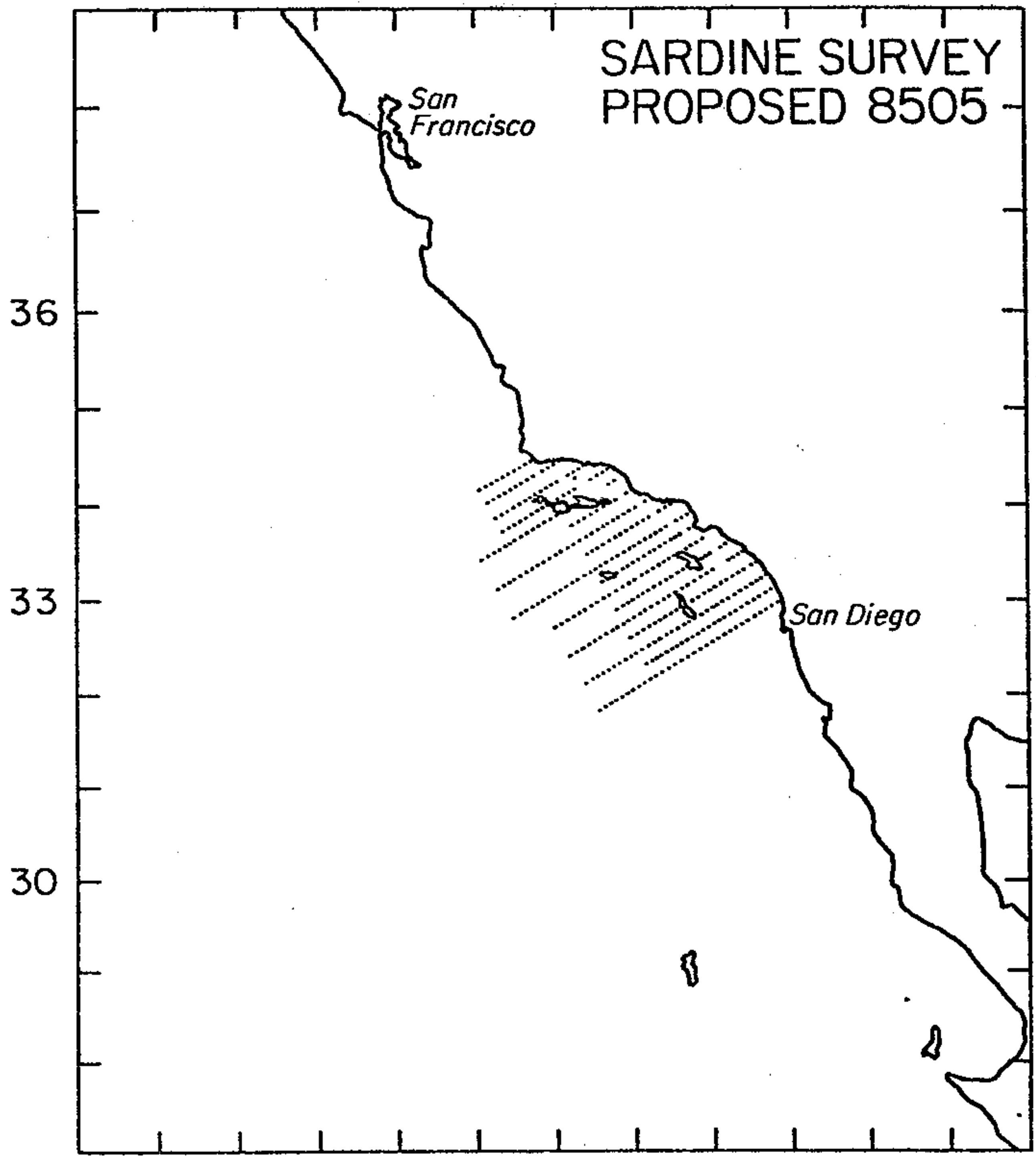
Figure 6. Three possible habitat models, p (S) (panels a) - c)) and the set-up for Bayesian analysis of the exit problem (panel d)).

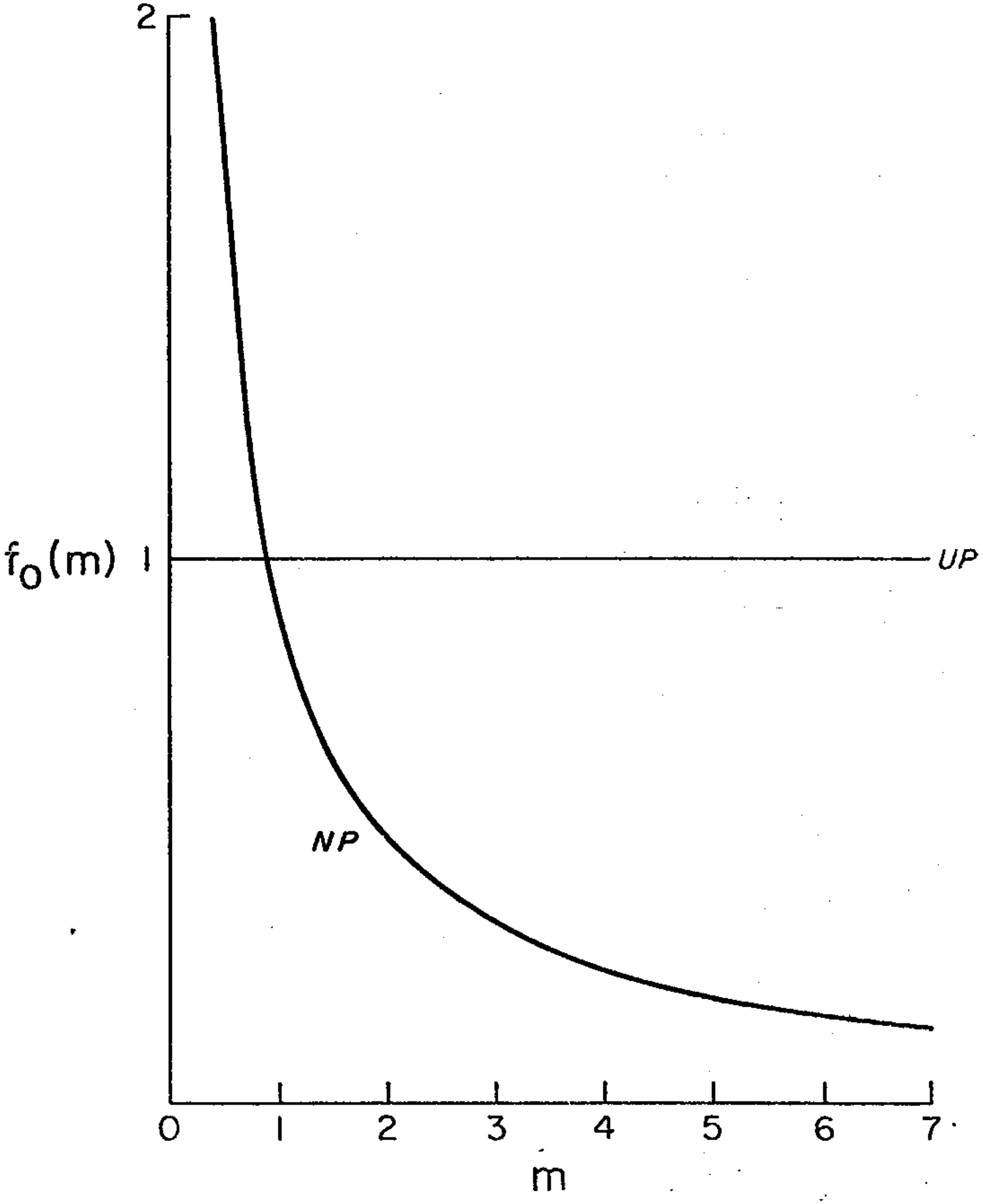
Figure A-1. Variance-mean ratio for the Z/R model.

Ring Fig 1

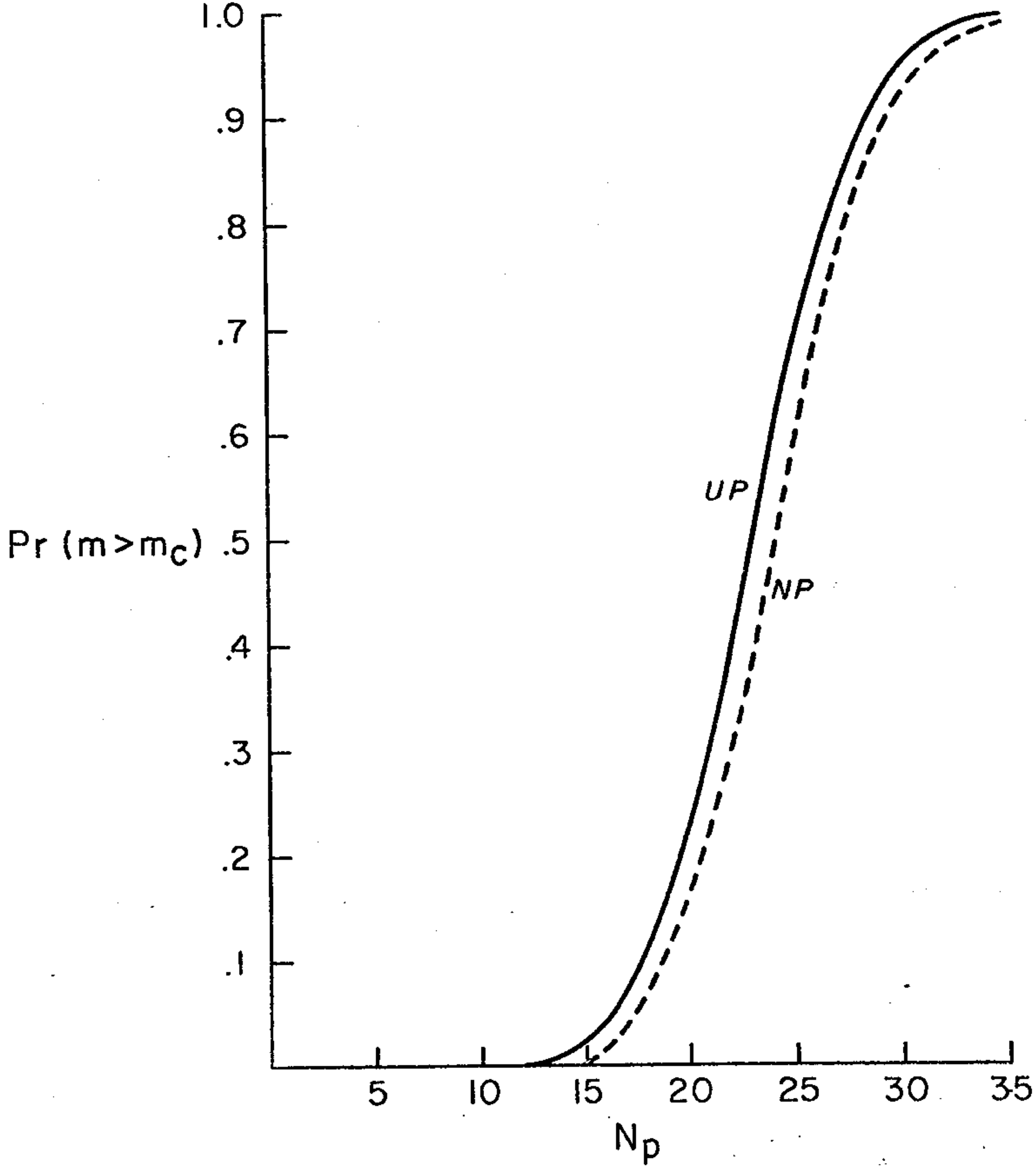


Mangel Fig 2

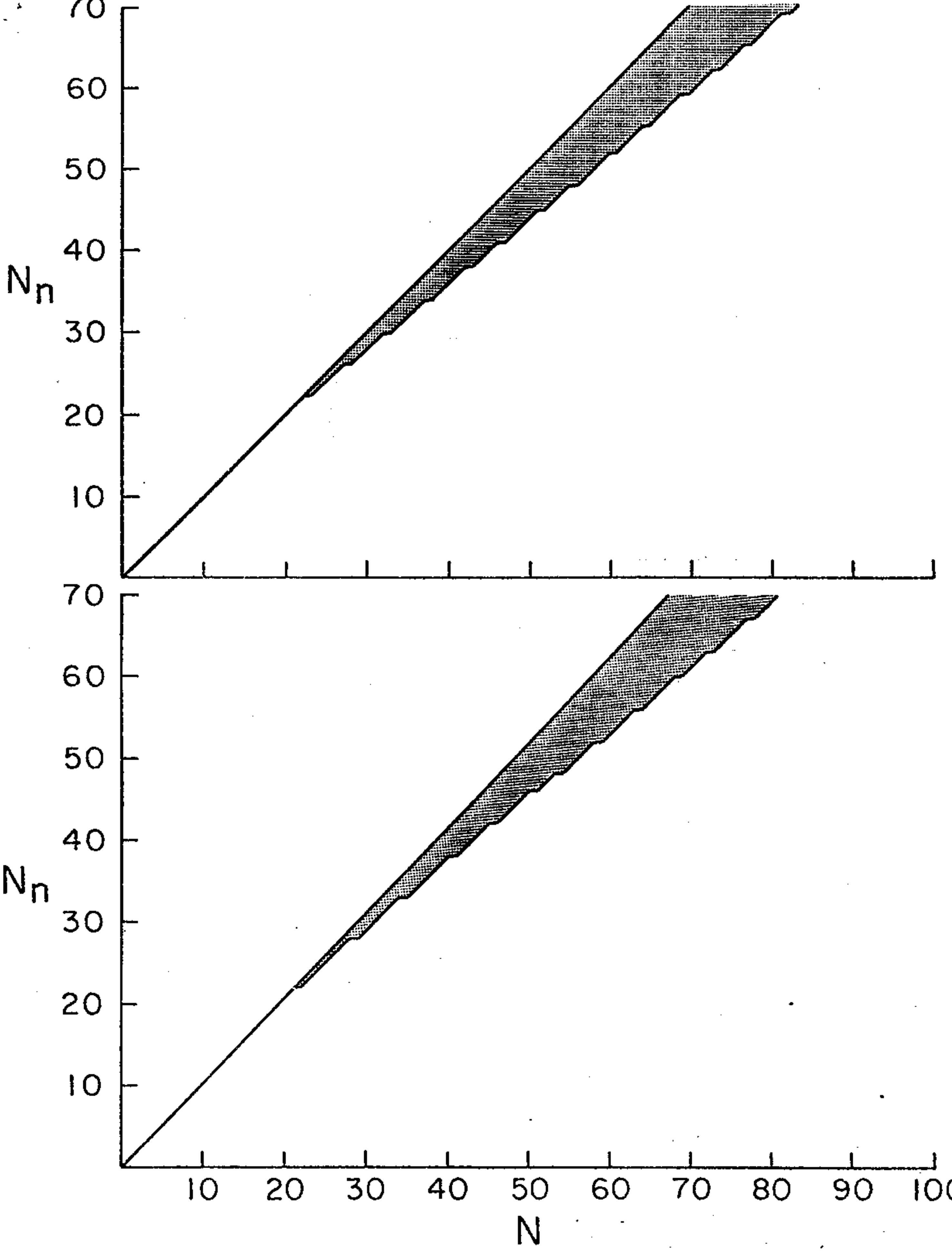

SARDINE SURVEY
PROPOSED 8505
San Francisco
San Diego
36
33
30

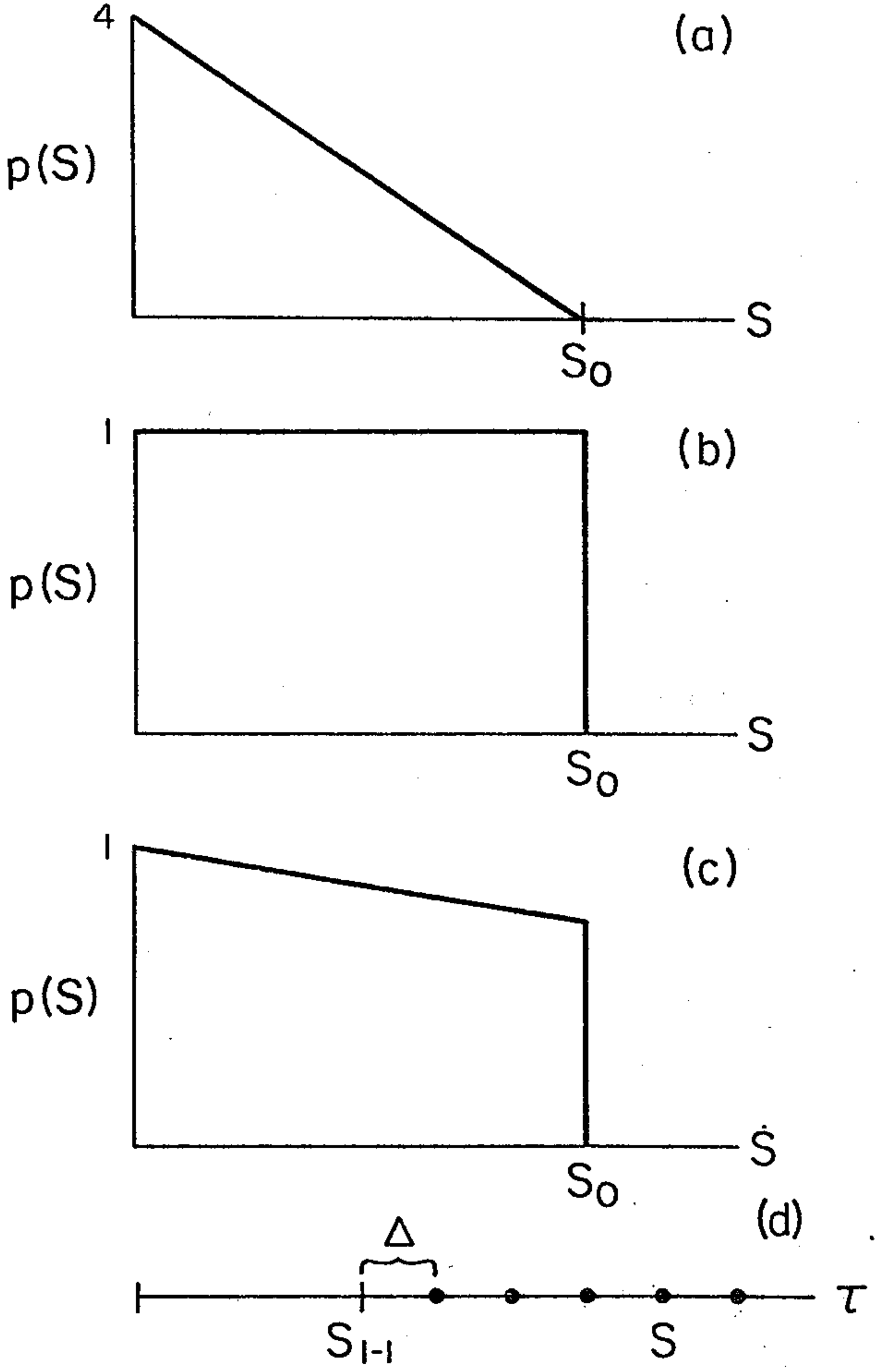

2
f_0(m)
1
UP
NP
0
1
2
3
4
5
6
7
m

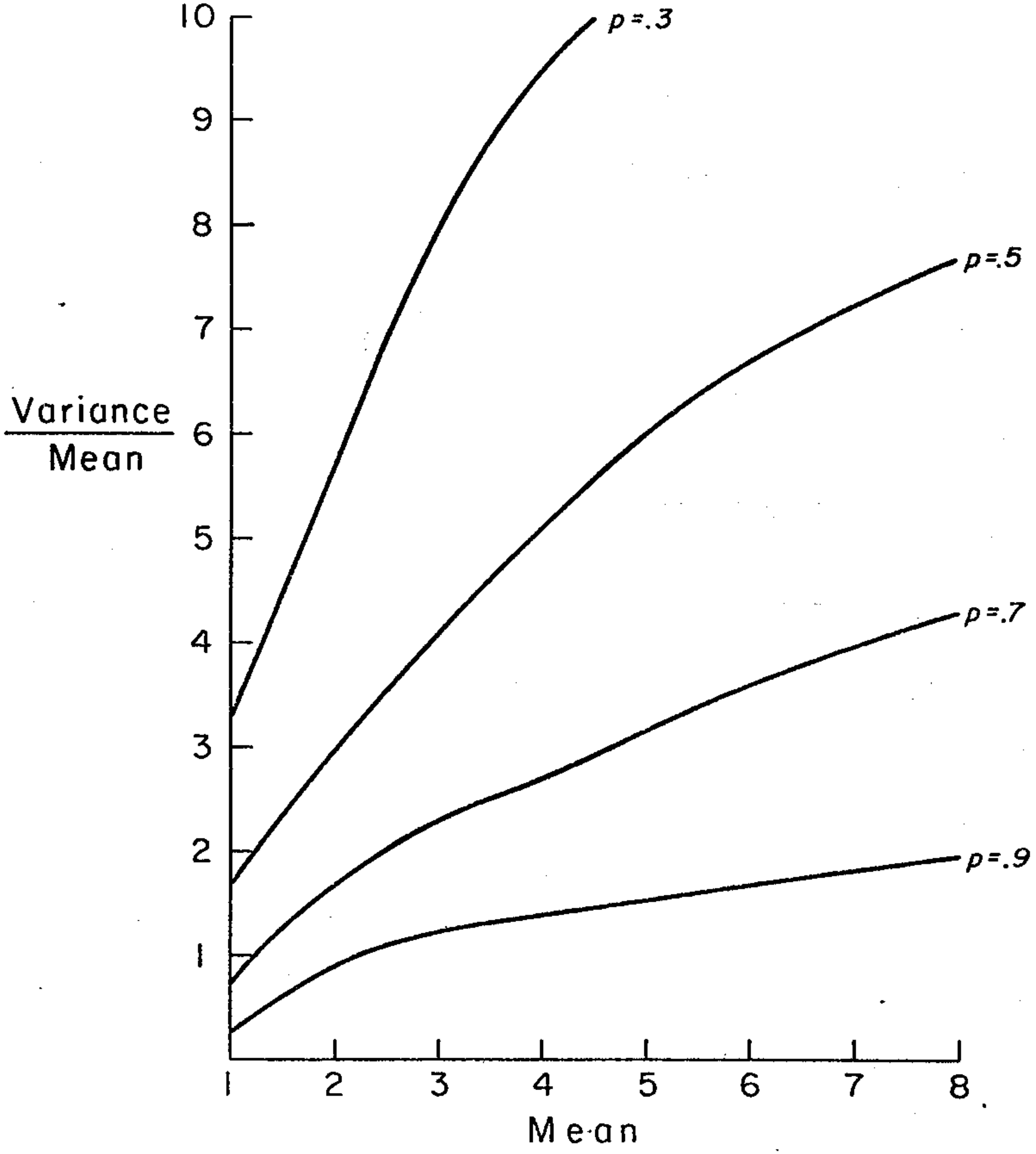
([illegible])



1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
Pr (m>m$_c$)
UP
NP
5
10
15
20
25
30
35
N$_p$


70
60
50
40
N_n
30
20
10
70
60
50
40
N_n
30
20
10
10
20
30
40
50
60
70
80
90
10C
N


4
(a)
p(S)
S
S₀
1
(b)
p(S)
S
S₀
1
(c)
p(S)
S
S₀
(d)
Δ
τ
S_{l-1}
S



p=.3
p=.5
p=.7
p=.9
10
9
8
7
Variance / Mean
6
5
4
3
2
1
1
2
3
4
5
6
7
8
Mean