Data-based Sampling and Model-based Estimation for Environmental Resources

G. P. Patil, G. J. Babu, R. C. Hennemuth, W. L. Myers, M. B. Rajarshi and C. Taillie

1. Introduction

Orthodox doctrine of sampling in applied statistics begins with the assumption of a well-defined population of interest, the study of which proceeds through acquisition of data having known probability of procurement defined by the sampling design. The key to this classical protocol for empirical extension of knowledge lies in the well-defined nature of the population and in the assigned probability of sampling its members.

Populations in nature, however, are defined by natural processes and one cannot quite construct the probability of encountering their members under arbitrarily assumed axiomatic behaviors. When populations are thus incompletely specified or incompletely accessible, it becomes necessary to use available data, and data that become available, for progressively more precise formulation and parameterization of population models. The population models, in turn, provide expectations concerning encounters that lead to estimators for parameters.

The purpose of the present paper is to provide insights into the approaches by which the interactions between data and design evolve into the enhanced knowledge of environmental attributes and ecological processes, while avoiding the cyclic problem of new information becoming non-information.

2. Data-based definitions of populations

Environmental investigations often begin at a stage where the kinds of entities that could comprise populations of interest and the extent of their occurrence are incompletely known. We begin instead with a universe of concern which may contain populations of interest. The universe of concern frequently has a strong spatial context, and can often be spatially delimited as a particular region of the biosphere. The potential populations of interest tend to be physical or biological
phenomena or objects. For example, possible occurrence of stress conditions in organisms may be of concern as an indication of environmental degradation—even in the absence of prior knowledge regarding the kinds and numbers of organisms present.

Pursuit of such environmental concerns typically begins with a search for existing data that might provide indications of variation within the universe of concern. There is, of course, the realization that encountered data probably include components of variance that are not of present interest. Nevertheless, possibilities that the variance structures reveal differences of interest usually make such data sets deserving of careful examination. Interest lies in statistical techniques that allow the investigator to determine which aspects of encountered variation are relevant (information) and which are extraneous (noise).

In a sense similar to the use of the term 'generic' in referring to drugs, we might think of generic data in the present context as being data likely to have utility for several types of problems in environmental analysis. For example, modern technologies have made remotely sensed data one of the most readily available generic sources of environmental information. The generic character of remotely sensed data arises from the fact that spectrally specific reflectance of environmental surfaces is largely determined by the composition and condition of those surfaces. The data are 'encountered' in the sense that there is seldom an opportunity to have the sensors custom configured in terms of spectral bands, resolution, and time of data collection. Remote sensing thus provides an illustration of spatially specific data that may be encountered by the environmental scientist concerned with determining possible occurrence and extent of phenomena that might indicate need for managerial or regulatory action. One strategy for statistical analysis puts the scientist precisely in the position of asking what population the sample represents.

2.1. A null hypothesis addressed in a meta-analytical mode

A simple null hypothesis can be posed as a point of beginning for analysis. It is that there are no differences in the environmental surfaces exposed to the sensor over the region of concern. Quantitative evidence bearing on the acceptance or rejection of this null hypothesis resides in spectral measurements made by the sensor. Since sensors usually make simultaneous measurements in several spectral bands, the data have a multivariate character. The decision to accept or reject the hypothesis is largely subjective, however, due to the lack of ability to construct a probability-based test criterion. Nevertheless, the rather wide array of quasi-statistical techniques encompassed by the term 'cluster analysis' provides an assessment of possibilities for arranging the spectral observations in groups that are 'well-separated' according to some index of discrimination. Since the grouping is done with the intent of creating groups or 'clusters' that are well-separated by subjective factors of choice, there is no probability basis for declaring that the degree of grouping so observed is meaningful. The spatial specificity of the spectral measurements coupled with the tendency of environmental phenomena to exhibit
spatial contagion provides a means of partially avoiding the specter of circular reasoning. Before pursuing this avenue of inquiry, however, it would be well to note the need for subsampling the remotely sensed data in connection with cluster analysis.

The resolution level of modern remote sensors provides a capability for generating millions of observations over areas of sizes routinely studied for purposes of environmental analysis. A moderate level of resolution, for instance, involves one observation for each 30-meter square on the environmental surface. The more sophisticated algorithms for cluster analysis typically require all possible comparisons between subgroups of observations, and several even require all possible comparisons over the entire set of observations. Even though remote sensors typically provide complete areal coverage in terms of observations, it would clearly be a monumental task for the largest computers to run such a data set through one of these algorithms in its entirety. Therefore, it becomes necessary to take a sample of the remotely sensed observations in order to make cluster analysis practical from a computational standpoint.

Another possibility for reducing computational burden lies in data compression relative to variates. It is commonly observed that measurements made in adjacent spectral bands are highly correlated. Principal component analysis serves to compress information onto fewer variates by segregating correlated aspects of variation. This often has a beneficial effect on the cluster analysis by removing implicit weighting arising from collinearity among variables. A further benefit is often realized from the action of principal component analysis as a noise filter.

For reasons indicated earlier, cluster analysis is more descriptive than inferential. Ecologists are well aware, however, that environmental phenomena tend to be contagiously distributed in a spatial sense. If the clustering could be made to encompass the entire set of spectral observations, then one might hope to distinguish environmentally induced aspects of clustering from artifacts of random variation according to spatial pattern. When the clusters are depicted on an image display device in correct spatial perspective, one would expect to find environmentally induced clusters occurring in a patchwork arrangement whereas artifacts of random variation should also be more or less random in their spatial distribution. We are reminded at this point that spatial patterns will be poorly expressed, if at all, when cluster assignments have been made for only a sparse sample of spectral observations.

In order to extend the assignment of cluster membership beyond the original sample for which clustering was performed, the analyst can treat the clusters as class prototypes and invoke statistical classification methods. A suite of such classification methods is available, differing primarily with respect to assumptions regarding distributional differences among classes. The less restrictive the assumptions, the more computationally complex (and consequently expensive) the algorithm. After extending the clustering through classification, the analyst is able to view an image-like representation of spatial distribution in an index of spectral non-uniformity over the environmental surface. If the display indicates a perceivable presence of spatial pattern over the environmental surface, the analyst
would reject the null hypothesis of environmental uniformity and would proceed to inquire regarding the causal basis for the perceived differences. It would be possible to conduct probability-based tests of nonrandomness in the spatial distribution of clusters by sampling the map through randomly located quadrats in parallel with the methods used by field ecologists. Perceived as discontinuities or intermittencies, spatial contagion is usually of such a degree as to be obvious. What we need is statistics to elaborate the not so obvious.

2.2. Sampling to identify populations

The foregoing scenario has presumably served to partition or stratify the universe of concern, with the partitions being relatively homogeneous in terms of ecological setting. The presumption of ecological homogeneity is, of course, based on similarity of spectral expression in the domain of the sensor. This presumption in turn becomes a null hypothesis of no differences within cluster unit to be tested objectively. Likewise, knowledge is lacking as to the specific ecological character of any given cluster unit. At this stage, then, the clusters become potential populations of interest, and the task becomes one of collecting data to determine the ecological identity of each cluster and decide whether it warrants further study.

In the terminology of sampling frames, each cluster becomes a stratum. With respect to sampling scenarios, the situation is essentially one of double sampling for (sub)stratification. The determination to be made on a sampled unit is a decision whether or not that unit is of further interest. If the proportion of sampled units in a stratum is less than some predetermined value, then the stratum as a whole will be deemed not to constitute a population of interest. In order to localize the effort involved in obtaining sample data for a stratum, it may be appropriate to treat spatially disjoint patches as primary sampling units in a multistage design.

It is possible that at this point the situation assumes the character of a normal sampling design problem in preparation for collection of new data to answer the questions posed. Alternatively, there may be a further appeal to existence of higher resolution remotely sensed data available from archives of agencies. Very possibly it will be in order to obtain photographic images from which objects can be identified visually by a human interpreter as opposed to computer-processed multispectral data (multivariate spectral measurements made by nonphotographic sensors).

Having located the populations of environmental interest in this manner, the investigator proceeds to focus on each of these populations individually and elaborate a data acquisition scheme for completion of the analysis. These data acquisition schemes may be, in turn, multiphase and/or multistage—possibly involving several different kinds of determinations made from encountered data. An important point is that encountered data should be incorporated in such a way that it does not constrain the capacity for ultimate inference.
3. Sampling design as design of encounters—The case of living marine resources

Study of marine ecology has been strongly motivated by the development and conservation of fisheries and the longstanding value to society of fish as food. The fishermen themselves sample the resource, using a variety of gear, and were the first to observe the inherent variability of the success of such encounters. They designed their encounters to maximize the catch per unit of fishing time and effort. The fisheries biologist’s first knowledge of the distribution and abundance of the resources came from observations of the fisherman’s practices and results.

The desire to better define the fluctuation in resource abundance and the reasons for it inevitably led to the conduct of sampling by the researchers themselves, which would overcome the recognized biases of the fisherman’s catch. The rapid post-World-War-II expansion of fisheries elevated the concern for conservation of the resources and management of fisheries, and led to increased resource ‘survey’ activity. The development of biometrics in agriculture was adopted as the basis for designing and analysis of the surveys. The same approach was also taken for studies of most other natural living resources such as forest and wildlife, but at least the degree, if not the fundamental nature, of difficulties of encountering and observing marine resources is different from that of observing terrestrial resources (Darr and Hennemuth, 1985).

The research surveys of marine resources cover all life stages—adults, eggs, larvae, and juveniles. Different types of sampling gear are used for each of the life stages and for different species, depending on the habitat occupied and the behavioral characteristics. For adults, the gear is mostly an adaptation, to some degree, of that used in commercial fisheries. For other life stages, the gear has usually been developed to meet variously defined standards of representativeness and consistency constrained by physical and logistical factors. Thus, the design of the sampling gear is the initial, and important, stage of designing encounters with the organisms, the characteristics of which we want to observe.

The second stage of design has been to determine how the gear is to be employed in time and space. The first and second stages described here involve some of the same factors, but the second stage is more related to the consideration of precision, whereas accuracy is taken to have been solved in the first stage—e.g., the size or species bias. The second stage is then concerned with selection of a statistical design of sampling and the associated data analysis and inference. Traditional developments in current statistical texts on design and analysis of surveys and experiments are usually applied.

These surveys have proven very useful in providing a fishery-independent measure of the relative magnitude and biological attributes of the stocks. The surveys have met the need for comprehensive qualitative descriptions; and the inclusion of species for which there is not a directed fishery and, hence, poor fishery statistics, has been particularly valuable. Use of the survey data for quantitative analysis has been more difficult; the validity of statistical inferences has frequently enough been questioned, and with some justification.
The areas sampled are large in most cases. The sampling fractions considered only in two dimensions (bottom or surface area) are often of the order of 10^{-6}. The relatively high variability of point estimates (coefficients of variation of 100% are not unusual), which seems characteristic of most survey data, is, perhaps, tolerable providing that it does reflect 'true' error variance. In other words, our observations are a representative sample for the hypothesized model. We are then imprecise, but can draw inferences that are at least vaguely right. What is certainly less tolerable is to be inaccurate, no matter how precise. This latter phenomenon is evident at times, and we can suitably qualify our conclusions, but even infrequent large and unexplainable deviations cast doubt on the soundness of our advice based on such data and the data analyses.

The process of 'designing' the surveys must, therefore, deal with the assumption of traditional, industrial statistics of design and analysis. We may need to have protocols for assuring consistency in our sampling gear and the way it is applied, so that bias due to this factor can be minimized. This will not, however, suffice for designing our surveys. A purpose of this section is to try to define what might suffice. This will be done in the context of encountered data in marine resource surveys. Improved ecological knowledge is really necessary for 'solving' the problem. In the mean time, there is a need for methods that deal with unpredictable and seemingly sporadic encounters.

Ancillary observations of the physical environment are made while sampling the resources, and direct surveys are also taken to provide information on environmental factors which might affect their distribution and availability. Very little attention is paid to the statistical distribution of these natural physical variables—e.g., temperature, water density, currents. Much more concern has been directed to determining concentrations and distribution of pollutants. Many of the same difficulties of encountered data apply to these aspects.

3.1. The issues

In marine ecology, the primary issue is how to develop an appropriate basis for drawing valid conclusions from the observations we have obtained. This applies to experiments, to surveys, and to monitoring, nearly equally. Of principal concern is accuracy or bias; in the context of this article, precision is involved to the extent that eliminating bias enables one to estimate error variance correctly.

In marine ecology, we are faced with a very dynamic physical and biological system over which the scientist has no means of control. All observations are taken under a different set of state conditions. For the most part, the targets are moving in an opaque fluid. The system may be said to be variable only in the sense that weather and people's reaction to it are said to be variable, i.e., there are physical and biological laws operating, but the multiplicity of causal factors and interactions make the system appear stochastic.

The unexplained variance—really differences between what we observe and what we might expect cannot be incorporated in a stochastic variable in the sense of 'noise'. This would be as great a mistake as assuming that an observation came
from a set of replicable events. With attribute sampling, a binomial error term (or trinomial) is assumed. Often, however, there remains an 'extraneous' portion of variation not accounted for by the model (Moore, 1987). The extraneous variation has been found to be upwards of 70% in samples of length classes.

Thus, the factors which affect catchability (or sightability, etc.) must be incorporated in the statistical analysis in a manner which distinguishes between accuracy and precision.

Most living organisms are found in clusters; i.e., they have a spatial dimension which results from some fundamental behavior mechanisms reacting to environmental factors. Statistical distributions were developed to express the behavior of variables in relation to an artificial set of postulates, and they cannot be expected to describe spatial attributes of organisms.

Our observations of an organism's density are a consequence of the encounter of the sampling gear with the aggregations of the organism. We may capture all or part of the aggregation depending on the relative size of the gear to the aggregation, or all or parts of several aggregations. The aggregation, and aggregations of aggregations, change their characteristics with changing activities—feeding, spawning, etc.—and changing magnitudes of the population. Environmental physical factors will also affect them.

If our 'sampler' is large relative to the aggregations, there may be a tendency to approach a (0, 1) variable (i.e., we are likely to encounter all of the aggregation or none at all). At another extreme, say with fewer, very large aggregations, the same sampler, now relatively smaller, would tend to generate a more continuous variable by including parts of the larger aggregations. The possibilities are infinite. The tendency is generally to fix the sampler (and method of application), thus we cannot detect such changes.

Much of the survey activity leads to a time series, from which inferences about the change on some process are drawn. Seasonal and annual series are common. It is often impossible to 'randomize' observations on the time series and the alternative is to fix a time of observation relative to some attribute of the environment or activity of the organism.

The natural—and now increasingly man-induced—events that effect real changes in populations and also the accuracy of observations may indeed be repetitious, but operate on various time and space scales in varying degrees. Thus, the seasons always come and go, and within less than epochal time scale, it will be warmer in summer than winter. However, each month, year, and decade will likely be different enough so that the biological response will be significantly altered.

The statistical issue is how to analyze the data when fixed samples are taken from a variable event. The question of what is a random independent sample in these circumstances seems moot. How can valid statistical inferences be drawn?

Thus, the sampling and analysis must begin from the perspective that the effect we are trying to estimate is not observed randomly over replicate conditions. Instead, there is a nested set of repetition of events, each varying with its own pattern that is, by and large, unpredictable. That does not mean that the condi-
tions or state of the system is not foreordained. We are observing the results of both the biological and physical processes, and their interactions—i.e., ecology. Under these circumstances, the traditional concepts of error, mean square, bias, and randomness, must be modified to deal with the reality. This comes down to correct identification of expectations, and using statistical models and analysis which fit the observed conditions rather than the other way around.

3.2. Some examples

Most marine organisms have a cyclical pattern of reproduction. The period may be several cycles per year for phytoplankton; a more continuous activity through the year for tropical fish; or shorter-term, distinct, spawning, egg, and larval stages within a year for polar and temperate fish.

A variable of great interest is the number of survivors of an annual spawning which 'recruit'—i.e., become available—to the fishery. The annual recruitment determines to a great extent the possible fishery yields over the near surface. There are two aspects of particular interest (Hennemuth, 1979):

First, the short-term advice to fishery managers is based on estimates of the fishable biomass which depend on the recruitment estimates of mortality, and gain in fish weight. In the longer run, determining the causal factors for success of recruitment, particularly the effect of spawning stock, has been a traditional concern. Lacking knowledge of cause of yearly differences in recruitment, the time series may provide knowledge of the expected year-to-year changes and longer-run consequences in a probabilistic context (Hennemuth and Avtges, 1982).

One gets only one observation per year, however, and the number of observations in the time-series is limited—generally well below which is dear to a statistician's heart. Further, the multiple factors which cause each year's results are difficult to observe, and replicability of them is an unrealizable concept.

One attempt to provide a valid statistical analysis is described in Section 6.

Seasonal trawl surveys (inventories) of fish have been conducted off the northeast coast of the U.S. for many years. These are stratified random trawl hauls. The operational limitations of one vessel over 72,000 n.m.² reduces spatial and temporal synopticity relative to the more remote sensing which introduces this article. The spatial density distribution of the fish is unknown in advance; in fact, it is unknown in general, except that it is 'contagious'. Doubtless, we are faced with nested distributions, in the 'burst' context described by Mandelbrot (1977). The aggregates are the result of the biological behavior—pairs at spawning, aggregation of pairs in physically optimal locales, aggregations of these aggregations over the species-preferred regions. Furthermore, the spatial density distribution is a variable itself, depending on environmental conditions that change in time and space.

Catchability of the trawl for each species, the age (life stage) of the fish, and activity of the fish—e.g., feeding, spawning, etc.—different. This is generally a seasonally changing phenomenon.

The season has been defined by conducting the survey within a fixed time, e.g., October and November in the 'autumn'. One knows, however, that the 'seasonal'
factors which control the biological behavioral responses of the fish are different from year to year, and predictable only within some time span which would encompass the whole range of possible seasonality.

Thus, there is some 'probability' distribution (a time series) of the time of the start of a season, and another probability distribution of the rate of change of events and their duration. One possible approach to designing our analysis, if not encounters, would be to incorporate these distributions.

An interesting consequence of seasonal variability is illustrated by the Georges Bank haddock. Schuck (1949) had used the observed seasonal (intra-annual) decrease in commercial catch per unit effort to estimate the actual weight of a year class removed by fishing. Subsequent analysis of survey catches indicated, however, that the catchability of haddock changed seasonally, and that this change was different for different age groups (Hennemuth et al., 1985). Thus, a significant part of the observed intra-annual decrease in catch per unit effort was due to catchability and not the removal of fish from the population.

Further analysis has indicated that intra-annual, seasonal variation in catchability is not consistent from year to year. This would preclude combining seasonal surveys unless additional information about causal factors can be obtained.

It is important to consider that the probability density functions which represent instantaneous spatial distributions contain variables (as parameters or functions) to accommodate changes through time, in environment or population states. This may be in the obvious-but-impossible category.

4. Sampling design as design of encounters—The bias in fisheries harvest data

4.1. Introduction

Non-response is a rather common feature in surveys of human subjects. One of the usual assumptions for addressing the problem of non-response, is that the characteristic of the population members on which information is being sought, does not differ between the groups of respondents and non-respondents. This allows us to regard the sample of those who respond as a random sample (of reduced size) from the entire population. Another problem which sample survey scientists often face is the possibility of getting a biased or inaccurate response. The combined effect of these two factors is that the mean of the numerical characteristic under investigation cannot be estimated by standard methods.

Indeed, we need to model characteristics that make a population member respond and/or give a biased response (if selected in the sample). What we observe is not a random sample. The event of response itself has a chance mechanism that may be related to the quantity being investigated. Statistical scientists need to modify both the interpretation of the sampling design and the estimation strategies. Following is an account of such an attempt in a survey of fishermen who report their commercial catch and data related to fishing efforts.
The Maryland Department of Natural Resources and the Virginia Marine Resources Commission conduct sample surveys to estimate the total catch of commercial as well as recreational fishery. These estimates find a number of important applications in resource management: providing indices of species abundance; documenting economic value; modelling fish population dynamics; etc. See Bonzek, Myers, Parolari and Patil (1986).

Agencies in charge of the surveys send mail questionnaires to a randomly selected sample of fishermen on a periodic basis. The samples are drawn by following a stratified sampling design, strata being based on license type and counties of residence of fishermen, so that no fisherman receives the questionnaire in two consecutive months.

A rather crude picture, that emerges from a study of the surveys over the years is that a number of fishermen do not respond at all (the non-response rate varies from 20% to 40%), and that among those who report, some report zero catch. Under-reporting is also suspected.

The commercial fishery survey offers two rather significant departures from the standard survey approaches. These amount to regarding the sample of fishermen who respond as a random sample (of a reduced size) from the target population. The fishing pattern of non-respondents may be dramatically different from that of the respondents.

Under-reporting by fishermen who report may have to do with an imagined connection with the federal and state government agencies responsible for regulation and revenue. Even when the surveying agency assures confidentiality of reported catches with the associated information on fishing effort, an application of randomized response survey may be in order.

Warner (1965), in a pioneering paper, introduced randomized response surveys to increase the response rate in surveys which involve questions that are sensitive either because of a social-religious stigma attached to them or because of possible legal implications of the response. Warner’s technique can be illustrated as follows. Suppose we are interested in estimating the proportion of persons who submitted an incorrect tax return last year. The randomized response questionnaire has two questions:

1. Did you submit an incorrect tax return last year?
2. Did you submit a correct tax return last year?

A respondent selects (1) with probability $p$ and (2) with probability $q = 1 - p$, $p$ being known to the survey management. It is expected that the respondent, once he selects a question randomly, responds truthfully, in view of the assurance of confidentiality (since the interviewer/management does not know the question the respondent is answering). The observed proportion of ‘yes’ answers in the sample is linearly related to the unknown proportion $\pi$ of those who submitted incorrect returns last year. This allows a numerical estimate of $\pi$.

Randomized response technique known as unrelated question design is due to Horvitz, Shah and Simmons (1967). Again, the respondent selects either of the two questions based on the outcome of a random experiment. However, only one of the questions is related to the sensitive character and the other is not. The true
proportion of 'yes' answers to the unrelated question would be either known or can be estimated from an independent survey. Unlike Warner's technique, the respondent has a feeling of assurance that he answered in a way, an absolutely unrelated question, resulting in more protection than Warner's technique.

Since Warner's introduction of the randomized response technique, there has been extensive work on statistical and logical aspects of randomized response techniques. (See Emrich (1983) and Fox and Tracy (1986). If the answer to the sensitive question is a numerical score, see Pollock and Beck (1976), Duffy and Waterton (1984), and others.) These techniques include extension of unrelated question design to the case of numerical response. In these methods, the respondent is asked to add or multiply the true score by a random number and report the final number.

Success of randomized response technique depends upon several factors including how the respondent views the randomization device, and assesses the degree of protection offered to him. Since the randomized response technique estimates have a larger standard error, we need larger sample size. Also, a respondent may not view the survey as a serious attempt to obtain relevant information because it incorporates something so very extraneous!

It appears that in these surveys, community interest and mistrust are at issue, and not a procedural guarantee of personal confidentiality. A statistical technique which assures the respondent of confidentiality or offers anonymity may be successful in sociological surveys on illegal abortion, drunken driving, child abuse, etc. But, if the respondent feels that the data gathered through such surveys are likely to be used against himself or his community, then the success of randomized response cannot be taken for granted.

5. Sample-based modeling of populations—An approach with weighted distributions

5.1. Introduction

The concept of weighted distributions can be traced to the study of the effects of methods of ascertainment upon estimation of frequencies by Fisher in 1934. It was formulated by Rao (1965) as a problem of specification of the model for observed data. Since then, weighted distributions have served as a very useful tool in the selection of an appropriate model for the observational process, especially when samples are drawn without a (proper) frame. See Patil and Rao (1977), Patil (1981), Rao (1985), Patil, Rao and Zelen (1987) and the cited references for interesting discussions on weighted distributions and their applications.

Let \( w(x, \omega) \geq 0 \) be a weight function, and let \( f(x; \theta) \) be a probability density function (pdf), such that

\[
\int w(x, \omega)f(x; \theta) \, dx < \infty.
\]
The distribution with pdf

\[ f^w(x; \alpha, \theta) = \frac{w(x, \alpha) f(x; \theta)}{E[w(X, \alpha)]} \quad (5.1.1) \]

is known as the weighted version of the distribution with pdf \( f(x; \theta) \). Weighted distributions corresponding to discrete distributions can be similarly defined. The weight function \( w(x, \alpha) \) is considered to be a factor explaining the probability by which the unit with value \( x \) enters the records/observational process. For example, the weight function \( w(x, \alpha) = x^\alpha \) with \( \alpha = \frac{1}{2} \) was found to give a good fit to data on the number of albino children in a family ascertained through affected children (Rao, 1965, 1985). The weight function \( w(x, \alpha) = x \) (i.e. \( \alpha = 1 \)) gives a distribution which is known as the size-biased distribution of \( X \). This weight function has been found applicable in many diverse areas of applications (Patil and Rao, 1977, 1978; Zelen and Feinleib, 1969).

In the above formulation, it is understood that the weight function does not depend upon the parameters of the random variable \( X \). In other words, \( w(x, \alpha) \) completely describes the properties of the observational process. Recently, we have discovered an application of weighted distributions where it is fruitful to extend the concept of the weight function by allowing it to depend upon \( \theta \), a parameter of the distribution of \( X \). Thus, the probability that a value \( x \) enters its data base depends upon the relative position of \( x \) vis-à-vis \( \theta \). As a result, we are able to extend the scope of weighted distributions where we model the observational process in relation to the parameter-related characteristics of the distribution. Models of overdispersion discussed by Diaconis and Efron (1985) and Efron (1986) provide an interesting application as briefly discussed below.

5.2. Double exponential family of distributions

Diaconis and Efron (1985) introduce a double exponential family of distributions in order to model overdispersion. Let

\[ f_{\mu, n}(x) = \exp \{n(\eta x - B(\mu) + C(x))\} = f(x) \quad (5.2.1) \]

be the pdf of a one-parameter exponential family of distributions with mean \( \mu \). Diaconis and Efron define the double exponential family corresponding to this distribution by

\[ g_{\mu, n}(x) = c(\mu, \theta, n) \theta^{1/2} [f_{\mu, n}(x)]^\theta [f_{\mu, n}(x)]^{1-\theta} = g(x) \quad \text{(say)} \]  

\( (5.2.2) \)

The parameter \( n \) in (5.2.1) and (5.2.2) is to be regarded as a sample size parameter. Rewrite (5.2.2) as

\[ g(x) = \frac{w(x, \mu, \theta) f(x)}{E[w(X, \mu, \theta)]} \quad (5.2.3) \]
where

\[ w(x, \mu, \theta) = \left[ \frac{f_{\mu, n}(x)}{f_{\mu, x}(x)} \right]^{1-\theta}. \] (5.2.4)

The distribution \( g(x) \) is a weighted distribution with a weight function depending on the parameter \( \mu \) of the original distribution.

We now discuss the implications of the weight function (5.2.4) in terms of the selection bias.

Notice that for the family \( f_{\mu, n}(x) \), \( x \) is the unique maximum likelihood estimator of the mean \( \mu \). Thus, if \( \theta < 1 \), \( w(x, \mu, \theta) > 1 \) for all \( x \). This offers an interesting interpretation of the density (5.2.2). Having realized \( x \), imagine that one carries out a likelihood ratio test for the hypothesis that the true mean is \( \mu \) and incorporates the observation with a probability proportional to weight function (5.2.4). That is, the observations which reject the true parameter value \( \mu \) are selected in the sample with a high probability; further, the smaller the \( p \)-values of these observations, the higher is their chance of being recorded in the sample.

More directly, this can be seen via \( I(x, \mu) \), the Kullback-Leibler distance function between \( x \) and \( \mu \). For, it is known that

\[ \frac{f_{x, n}(x)}{f_{\mu, n}(x)} = \exp \{ nI(x, \mu) \} \] (5.2.5)

(see Efron, 1986, p. 712); so that the weight function (5.2.4) reduces to

\[ w(x, \mu, \theta) = \exp \{ n(1-\theta)I(x, \mu) \}. \] (5.2.6)

Thus, the distribution (5.2.2) admits the values \( x \) in the sample space with probability proportional to their distance from the true parameter. The larger the distance, the larger is the chance of these observations being recorded. This explanation is entirely consistent with the usual understanding of the weighted distributions, in that the probability of recording an \( x \) is not the same for all \( x \) in the sample space. See also Royal and Cumberland (1981a, b) for some discussion on randomization and non-randomization.

In this connection, it is interesting to note that in his analysis of the data on disease toxoplasmosis in 34 cities of El Salvador, Efron (1986) argues that 'genuine random sampling was infeasible and the subjects may have been obtained in clumps'. In the data sets that he discusses, the frequency of very high counts as well as the frequency of very low counts is larger than expected when compared to the binomial and Poisson distributions. While carrying out the regression analysis of these counts on a covariate, such as the rainfall in his case, we need to attend to the selection bias. Efron's approach does this by bringing into the model a weight factor which accounts for possible non-random sampling.
6. Combining recruitment data and kernel approach

6.1. Background

For several years the Northeast Fisheries Center (NEFC) has been assembling recruitment series for a large number of oceanic fish stocks. Recruitment is defined by the number of fish of 'catchable' size entering a fish stock. Estimation of recruitment distributions is important for the assessment and prediction of long term frequencies of good and poor year classes. In this connection, several parametric distributional models have been fitted to each of the available recruitment data sets (Hennemuth, Palmer, and Brown, 1980; Patil and Taillie, 1981). The small sample sizes prevented reliable assessment of goodness-of-fit. It also proved difficult to effectively discriminate between competing models, e.g., between the gamma and the lognormal distribution.

In view of the above, it was suggested that the recruitment data for the various stocks be combined into a single large data set and analyzed with the two-fold purpose:

(i) to better assess the fitting performance of the different methods and models, and

(ii) to arrive at a fairly precise estimate for a 'universal' recruitment distribution.

6.2. Combining the data sets

Recruitment series for 18 stocks were selected for analysis. The data and histograms for the individual stocks appear in Figure 6.2.1. Sample sizes range from 10 for North Sea mackerel to 43 for Georges Bank haddock. On the whole, the data exhibit strong positive skewness with the occasional occurrence of large positive values corresponding to the appearance of a strong year class.

When combining data, the various data sets must have some common features (or there would be no reason to combine) as well as some differences (or the matter would be trivial). The trick is to model the common features and to suitably adjust the data for the differences before combining. The large combined data set is then used to draw reliable inferences concerning the common features.

In our case, it is hypothesized that the pth recruitment data set can be described as a random sample from a scale-parameter family of distributions with cumulative distribution function (cdf)

\[ F(x, \theta_p) \equiv F(x/\theta_p). \]  

(6.2.1)

Here the scale parameter \( \theta_p \) is allowed to vary from stock to stock. The functional form of the cdf \( F \) is assumed to be the same for all stocks and therefore represents a 'universal' recruitment distribution. The pth data set is adjusted by dividing through by a suitable scale statistic. The arithmetic mean (divided by 5) was used in the present analysis.
6.3. Estimating the universal recruitment distribution

Having combined the descaled recruitment values, the next step is estimation of the common cdf $F$. Here a nonparametric approach has been adopted. In passing, it may be noted that the problem would be trivial if we were prepared to assume a parametric form for $F$. In fact, if $F(\cdot) = G(\cdot, \phi)$ where $G$ is a known distribution and $\phi$ is a vector of unknown parameters, then from (6.2.1) the $p$th data set is a random sample from $G(x/\theta_p, \phi)$. From this, the joint likelihood can be written down and parameters estimated in the usual way.

The nonparametric method employed is a variation of the kernel technique. The distribution to be estimated is approximated by a mixture of lognormal distributions. There is one lognormal (known as a kernel) for each available observation. Each kernel is ‘centered’ so that its geometric mean is located at the corresponding observation. The various kernels are taken to have the same logarithmic standard deviation, which is known as the bandwidth.

The central theme in application of kernel methodology is the determination of suitable bandwidths. Overly small bandwidths yield estimated pdf’s whose graphs have a rough, jagged appearance. Excessively large bandwidths smooth the probability mass over a wide interval, losing most of the local features of the data. There is extensive literature on kernel methods (Wertz and Schneider, 1979; B. L. S. Prakasa Rao, 1983; Devroye and Gyorfi, 1985; Silverman, 1986) and we will not here dwell upon the technical aspects except to point out that bandwidth determination was done through cross-validation.

The histogram of the combined data set is shown in Figure 6.2.2. Superimposed are the fitted lognormal distribution, the fitted gamma distribution and the kernel estimate. Parameter estimation for the gamma and lognormal was done by the method of maximum likelihood. Inadequacy of the lognormal fit is readily apparent. The kernel fit, while generally acceptable, does exhibit a leftward bias for small year classes. We have been able to remove most of this bias by either of two techniques: (i) variable bandwidths and (ii) regression toward the mean of the kernel centers. The gamma distribution also shows a leftward bias, and has a right hand tail that is much too short to give an adequate fit.

6.4. Estimating individual recruitment distributions—Interpretation of the kernel estimator

The simplest estimate of the recruitment distribution for a particular stock is formed by rescaling the universal recruitment distribution. This raises the question of whether the limited data that is available on a particular stock can be used to improve the estimate. Here the James–Stein (1961) paradigm may offer some guidance. Envision the separate (descaled) recruitment distributions as forming a cloud of points in the space of all probability distributions. The universal curve estimates the center of this cloud. Use the available data to obtain, perhaps by the kernel method, a low quality estimate $\hat{F}_p$ for a particular distribution. For the final estimate, use a convex linear combination of the imprecise estimate $\hat{F}_p$ and the precise but inaccurate universal estimate.
Fig. 6.21. Recruitment data.
Fig. 6.2.1. (continued).
It may be of interest to close this section with an interpretation of the kernel estimator. The recruitment process is governed by many factors, both environmental and biological. Currently there is little understanding of what these factors are, how they operate quantitatively and how they interact. The kernel method attempts to account for the annual variability in recruitment without developing a detailed explanatory model. Consider the multidimensional space of all relevant factors and let this space be partitioned into \( N \) subsets, one for each available recruitment value; the subsets occur with the same long term relative frequency of \( 1/N \). Conditional upon a particular partitioning set, there is still residual environmental variability within that set and a corresponding variability in recruitment. It is this variability that is represented by the lognormal kernels. Each kernel is centered at the corresponding observation, in effect treating each observation as typical for its partition set.

7. Encountered ecotoxicological data and chronic effects thresholds

In problems related to fisheries, we often find ourselves interested in, say, two possible parameters \((L, M)\). To conduct experiments on a population of various species and to estimate \( M \) may be very expensive and time consuming, whereas, \( L \) may be obtained with comparatively little effort and cost. If there is a functional relation between \( M \) and \( L \), then \( M \) can be extrapolated for a given species from that of the estimated value of \( L \). In order to establish such a functional relationship, we need data on \((L, M)\). Due to the time and cost involved, ‘observations’ are ‘collected’ from experiments previously conducted in various laboratories. The data obtained do not constitute a random sample, and, therefore, standard statistical techniques, such as regression, cannot be applied. In this section, we discuss a statistical approach to deal with such problems and illustrate it with an example. For details, see Linder et al. (1986) and Suter et al. (1987).

Ecological effects of chemicals on commercial fish are commonly assessed by estimating a ‘safe’ exposure level, below which no observed effect on the growth, reproductivity and mortality of the fish occurs. These effects usually vary over species and toxicants. For a particular species-toxicant pair, the effects of chemical levels are significant at various life cycle stages of the species. The largest concentration level, below which no significant effect is observed, is estimated. This is a very expensive and time consuming process. It is not possible to conduct a test for every possible toxicant and species of interest. The proposed approach consists of looking at the following two kinds of effects.

1. **CHRONIC** Long term low-level effect; Maximum Acceptable Toxic Concentration (MATC), the largest concentration level below which no significant effect is observed.

2. **ACUTE** Short term high-level effect. Achieve 50% mortality in 96 hours (LC50).

If these paired results are available corresponding to several species-toxicant
pairs, a functional relationship, the so-called acute-chronic extrapolation, can be estimated. This relationship can then be used for extrapolation from LC50 to MATC.

This procedure involves curve fitting for the data points \((X_i, Y_i)\), the LC50–MATC pair for a particular toxicant species combination. The main features of the 'data set' are:

(i) Each point (or pair) represents reported results from two bio-assay experiments.

(ii) Different points result from different studies reported in literature. Hence, the \((X_i, Y_i)\) do not constitute a random sample from the population of all possible LC50–MATC pairs.

(iii) Since \((X_i, Y_i)\) are estimates of threshold concentrations, they are themselves random quantities. There is considerable uncertainty about their 'true' values.

Even though the pairs \((X_i, Y_i)\) are independent, the classical least squares method cannot be applied for curve fitting as the data is an encountered one, and with possible errors. It is not a random sample. This can be viewed as independent observations on the population \(\Omega = \{(s, t): s \text{ denotes species and } t \text{ denotes toxicant}\) for each \(\omega\) in \(\Omega\), let \((L(\omega), M(\omega))\) denote the actual values of LC50 and MATC. What we observe, for some of these species, is \(X = L + \text{error} \) and \(Y = M + \text{error}\). This leads us to the errors-in-variables model.

### 7.1. The errors-in-variables model

In the errors-in-variables (EIV) model, \((X_i, Y_i)\) are assumed to have been recorded with error \((\delta_i, \epsilon_i)\). They represent unknown mathematical quantities \((U_i, V_i)\). Linearity is assumed between \(U_i\) and \(V_i\), resulting in the model:

\[
X_i = U_i + \delta_i, \quad Y_i = V_i + \epsilon_i, \quad V_i = \alpha + \beta U_i
\]  \hspace{1cm} (7.1.1)

and

\[
(\delta_i, \epsilon_i) \sim \text{i.i.d.} \begin{pmatrix} 0, & (\sigma_\delta^2 & 0) \\ 0 & \sigma_\epsilon^2 \end{pmatrix}.
\]  \hspace{1cm} (7.1.2)

The \(\{\delta_i, \epsilon_i; i, j\}\) are independent identically distributed random variables with zero means and with \(\sigma_\delta^2\) and \(\sigma_\epsilon^2\) as variances.

Two structures are possible.

(i) Structural EIV model; and (ii) Functional EIV model

In (i) \(U_i\) are assumed to be independent identically distributed from an unknown distribution with mean \(\mu\) and variance \(\sigma^2\), and are independent of \(\{(\delta_i, \epsilon_i)\}\). In this case, \((X_i, Y_i)\) is a sequence of independent identically distributed vectors with mean \((\mu, \alpha + \beta \mu)\) and dispersion matrix

\[
\begin{pmatrix}
\sigma^2 + \sigma_\delta^2 & \beta \sigma^2 \\
\beta \sigma^2 & \beta^2 \sigma^2 + \sigma_\epsilon^2
\end{pmatrix}.
\]

The classical asymptotic theory is applicable for the structural model. But the
more interesting model in the present context is model (ii), where the \( U_1, \ldots, U_n \) are treated as unknown nuisance parameters. These two EIV models have been studied extensively (Kendall and Stuart, 1979; Gleser, 1983, 1985). For purposes of identifiability, both models assume \( \lambda = \sigma^2 / \sigma^2_\delta \) to be known.

The maximum likelihood estimators of the slope \( \beta \) and the intercept \( \alpha \) (under normality assumptions) for both the structural and the functional models are:

\[
\tilde{\beta} = h + \text{sign}(SXY) \sqrt{(h^2 + \lambda)}, \quad \hat{\alpha} = \bar{Y} - \tilde{\beta} \bar{X},
\]

(7.1.3)

where \( SXY = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) \) and \( h = (SYY - \lambda SXX) / 2SXY \),

\[
\hat{\sigma}_\delta^2 = \frac{1}{(\lambda + \hat{\beta}^2)} \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta}X_i)^2
\]

where bars denote averages and where SXX, and SYY are the usual corrected sums of squares.

7.2. Simulation results and data analysis

The following table gives the simulation results to compare OLS and Structural EIV models:

<table>
<thead>
<tr>
<th>True values ( \alpha = 0, \beta = 0.5, \lambda = \sigma_\epsilon = \sigma_\delta = 1 ); 20 observations, 30 simulations;</th>
<th>Average of OLS ( \hat{\beta} )</th>
<th>Average of EIV ( \tilde{\beta} )</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.479</td>
<td>0.501</td>
<td>( U_i \sim N(0, 25) )</td>
<td></td>
</tr>
<tr>
<td>0.489</td>
<td>0.500</td>
<td>( U_i \sim U(10, 10) )</td>
<td></td>
</tr>
</tbody>
</table>

For the data described above, a more realistic model may be a modified version of the functional EIV model with \( (\delta, \epsilon) \) independent but not identically distributed. Further, it may not be unrealistic to assume that \( (\delta, \epsilon) \) have the same dispersion matrix. Without loss of generality, assume that \( E(\delta) = E(\epsilon) = 0 \). (If they are not zero, merge these values in \( U_i \) and \( V_i \).) It can be shown, under these assumptions that, if \( \mu_\epsilon = (1/n) \sum U_i \) converges to \( \mu \) and if \( \sigma^2_\epsilon = (1/n) \sum (U_i - \mu_\epsilon)^2 \) converges to \( \sigma^2 > 0 \), then \( \tilde{\beta} \) and \( \hat{\alpha} \) are asymptotically consistent. Using the Lindeberg’s theorem and the so-called \( \delta \)-method we can show that \( \sqrt{n} (\tilde{\beta} - \beta) \) is asymptotically normal.

For the applicability of the EIV method, the results (absolute magnitudes) have to be evaluated in terms of their biological meaning. This has been done extensively (Suter et al., 1986) by comparing extrapolated MATC’s with measured MATC’s for toxicant-species combinations, where the results are available. In addition, the method was compared to the traditional methods which use fathead minnow as a basis for extrapolation. It has been generally found to out perform the old methods.
Often the sampling distributions of the estimates of \( \alpha \) and \( \beta \) are skewed. So the normal approximation is not reliable for moderate size samples, as it ignores the skewness factor. This is where the resampling method like bootstrap helps. It turns out that the bootstrap is not straightforward in the case of EIV model. The next section explains the bootstrap method for the EIV model.

### 7.3. The bootstrap method

Babu (1984) and Babu and Singh (1984) discuss the asymptotic properties of bootstrap. In order to use the bootstrap method, we need to 'imitate' the random structures involved. This amounts to choosing proper fitted values \( \hat{U}_i \) and \( \hat{V}_i \) so that the average \( |X_i - \hat{U}_i| \) is close to \( |Y_i - \hat{V}_i| \sqrt{\lambda} \). It turns out that the maximum likelihood estimates do not satisfy this restriction under normal assumptions. The adjusted residuals are given by

\[
d_i = -\frac{C_i \text{sign}(\hat{\beta})}{\sqrt{\lambda + \hat{\beta}^2}} \quad \text{and} \quad e_i = -\sqrt{\lambda} d_i \text{sign}(\hat{\beta}),
\]

where \( C_i = Y_i - \hat{\alpha} - \hat{\beta} X_i \) with \( \hat{U}_i = X_i + C_i/(\hat{\beta} + \sqrt{\lambda} \text{sign}(\hat{\beta})) \) and \( \hat{V}_i = \hat{\alpha} + \hat{\beta} \hat{U}_i \).

If we sample with replacement from \((d_i, e_i)\) in order to use the bootstrap method, the structure of the resulting residuals does not match that of \((\delta_i, \varepsilon_i)\). Further modification of the bootstrap sampling procedure is required. Need for such modifications is also pointed out in the case of autoregressive processes (see Babu and Bose, 1986).

The asymptotic theory plays an important role in suggesting the necessary adjustments and refinements for the bootstrap method. Thus let \( r_1^*, \ldots, r_n^* \) be a sample of size \( 2n \) from \( d_1, \ldots, d_n \) with replacement. The sample \( \{r_i^*, -\sqrt{\lambda} r_i^* \} \) behaves roughly like a sample with replacement from the unknown quantities \( (\delta_1, \varepsilon_1), \ldots, (\delta_n, \varepsilon_n) \). Then,

\[
X_i^* = \hat{U}_i + r_i^* \quad \text{and} \quad Y_i^* = \hat{V}_i - \sqrt{\lambda} r_i^* \]

have essentially the same structure as those of (7.1.1) and (7.1.2). Since \( \beta^* \) is computed from \((X^*, Y^*)\) as \( \hat{\beta} \) is from \((X, Y)\), as \( n \to \infty \),

\[
\sup_s |P^*(\sqrt{n}(\beta^* - \hat{\beta}) < xs^*) - P(\sqrt{n}(\hat{\beta} - \beta) < xs)| \to 0
\]

for almost all sample sequences, where \( s^* \) and \( s \) are the estimated standard deviations of \( \beta^* - \hat{\beta} \) and \( \hat{\beta} - \beta \) respectively. Even though no closed form expression for the distribution of \( \hat{\beta} \) is known, the bootstrap method gives a good approximation for the distribution of \( \hat{\beta} \), even in the non-normal case. Using Edgeworth expansions, we can show that the convergence here is faster than the convergence of the normal approximation.
8. Synthesis

As in the foregoing examples, it is common that classical theory of sampling cannot be directly applied in situations calling for quantification of environmental resources. Definitions of populations, and uncertainties relative to the stochastics of encounters, constitute fundamental problems with respect to the application of classical methodology in sample design and estimation. It becomes necessary to use existing data bases in conjunction with descriptive models of behaviors and processes to arrive at a workable, it not always exact, approach to further quantification.

Utilization of existing data may be considered as the ultimate lack of ability to design encounters. The contents of the data base have already been collected, so there is no design opportunity whatever. The data can only be subjected to (perhaps secondary) analysis. The focus turns to exploiting any ordinations common to both the data base and the field environment. Ordinations, such as space and time, carry universal relevance because of ecological linkages causing extensive covariation of environmental variables. The more the variables in a data base share a pattern of covariation in space and time, the more compelling the conjecture that the environmental spectrum extends to other variables. In other words, such covariance provides evidence of environmental regionalization. Such regions provide a natural basis for stratification, and the dimensions of ordination may provide a basis for a future sampling frame.

The pervasiveness of covariation in the spatial context among environmental resources serves to underscore the need for advances in multivariate techniques of regionalization. One can envision algorithmic examination of spatially specific environmental data bases for automated delineation of uniform regions to serve as strata. Likewise, cases where environmental variables recorded in a data base exhibit lack of coherence in space and time, indicate need for intensive data collection to determine causes of inconsistency. The latter are likely to be ecologically complex environmental settings and/or instances of alteration. In either case, such areas are of central concern for environmental monitoring.

If one cannot know the probability of observing a potential sampling unit, the next best thing is to have a means of modeling the probability. Knowledge of behavioral patterns and processes, whether or not formalized, provides a basis for modeling the probabilities of encounters.

Finally, the commonalities among populations may be exploited for purposes of combined estimation and extrapolation. Some approaches of this kind are among the illustrative examples.

Acknowledgements

Preparation of this paper has been partially supported by the NOAA research grants to the Penn State Center for Statistical Ecology and Environmental Statistics under the auspices of the Northeast Fisheries Center, the Chesapeake Bay
Stock Assessment Committee, and the Oceans Assessments Division. The NCI grant CA 23415 to Dana-Farber Cancer Institute provided partial support to G. P. Patil. Our thanks are due to Marilyn Boswell, Ernst Linder, and Joel O'Connor for interesting discussions on the contents of this paper from time to time.

References


