Encountered Data, Statistical Ecology, Environmental Statistics, and Weighted Distribution Methods

Ganapati P. Patil

ABSTRACT

We have begun to experience in data gathering and analysis in modern ecological and environmental work a space age/stone age syndrome. Also we are challenged to break into the cycle of no information, new information, and non-information while dealing with soft data, hard looks, and prudent decision-making involving errors of the 'third' and the 'fourth' kind in addition to those of the first and the second type.

Weighted distribution methods arise in the context of data gathering, modeling, inference, and computing, and help provide a unified approach in dealing with encountered data.

KEY WORDS: Double exponential family; Fisher information; Harvested equilibrium distribution; Length-biased sampling; Meta-analysis; Unreasonableness of environmental degradation.

1. STATISTICAL ECOLOGY, ENCOUNTERED DATA, AND META ANALYSIS

First of all, we must recognize that there is no substitute for good data, and in no way should we be complacent about it. As we know, statistical

---

1 Distinguished Professor of Mathematical Statistics and Director of the Center for Statistical Ecology and Environmental Statistics in the Department of Statistics at The Pennsylvania State University, University Park, PA 16802. This paper is based on the Invited General Methodology Lecture given by the author at the Joint Annual Statistics Meetings in San Francisco, August 1987 at which time he was also a Visiting Professor of Biostatistics in the Department of Biostatistics at Harvard University.
thinking is an aid to the interpretation of data. It may help clarify seeming confusion. It may help confuse seeming clarity. What is important is that the statistical approach is expected to contribute to the overall insight and perspective of the substantive issue and its resolution in the light of the evidence on hand, be it in the nature of empirical data, be it in the nature of literature assembled data, be it in the nature of expert opinion data or be it a combination of all three.

What makes the problem of ecological and environmental investigations different from studies in the physical sciences and in engineering is that, unlike in the physical sciences and engineering, we often have a longer span of investigation depending on life stages and age lengths. Also, there are instrumentation changes that are a consequence of advancing technology. That puts us in a rather difficult cycle of no information, new information and non-information. Our response to an ecological or environmental question usually is that we do not have enough information. We need to collect new information. And, after the new information is collected, we often must say that the information collected is already non-information. However, non-information is no information, and thus we go through a cycle involving data that does not produce information. We need to break this unaffordable cycle.

At the very least, we need to be aware of this fundamental issue in ecological and environmental work. It may be characterized by the concept of soft data. Not only is there biological or ecological or environmental variation contributing to large variability usually present in these data sets, but also there are measurement errors arising from a variety of sources — some known, some unknown, and some unknowable. Thus the data are soft data, and they call for not merely hard analysis, for which we have software and software packages, but they need very hard looks if we are expected to make prudent decisions. This is particularly true when we recognize that, while there are questionable statistical routines, there are no routine statistical questions.

Surely, each of us has cherished the hope that we might acquire a crystal ball for forecasting purposes; but we know that there are no crystal balls around. We can be happy if we deliver crystal cubes — a concept discussed in section 8! And when we speak of these crystal cubes, we are faced with a space age/stone age situation; space age data and space age analysis? Stone age data and stone age analysis? This is the dilemma we face — whether we find ourselves in the space age or the stone age, and whether we match or do not match.

We are left with comprehensive data. Does that mean that there is comprehensive information? And if there is comprehensive information, does that mean there are coherent conclusions? These problems are present in ecological and environmental work because of various problems with combining data involving heterogeneity, variability, uncertainty, or encountering.
And, encountering is one very common feature of ecological data. The investigator goes into the field, observes and records what he observes. Thus, the situation is not as in socioeconomic surveys where there is a sampling frame with sampling (selection) probabilities defining a regular sampling design. Rather here the investigator goes out and records what he encounters, and the data are by and large a combination of the encounter data types. This element is what both ecologists and the statisticians have to incorporate into their thinking, analysis and interpretation. The conclusions may be weak, but we know what it means when we are in the dark where only a few rays of light can mean so much. This is also true of sampling, modeling, forecasting, and management.

2. Certain Scientific Perspectives Through Examples

While progress within a single discipline is not necessarily easy, progress with interdisciplinary research and training tend to be even more difficult. The following stories may help convey some of the flavor and the essence of an interdisciplinary effort.

2.1 Comprehensive versus Comprehensible

Consider the following. We wish to comprehend a given situation.

- For lack of information, we do not (quite) comprehend the situation.
- We collect information, and we tend to collect comprehensive information.
- Because the information is comprehensive, we do not (quite) comprehend it.
- So we summarize the information through a set of indices (statistics) so that it should become comprehensible.
- But now, we do not comprehend quite what the indices exactly mean.
- And, therefore, we do not (quite) comprehend the situation.
- Thus, without (all) information, or with (partial) information, or with summarized information, we do not quite comprehend a situation!

This dilemma is not to suggest a bleak picture of one’s ability to understand, predict, or manage a situation in the face of uncertainty. It is more to suggest a need to state clearly the purpose, formulation, and solution for the study under consideration.
2.2 How Many of Them Are Out There?

This scenario takes place in a court of law. The issue is about the abundance of species seemingly endangered, threatened, or rare. The judge orders an investigation. A seasoned investigator conducts the survey. He reports having seen 75 individual members of the species under consideration. The judge invites comments.

**INDUSTRIAL LOBBY:** The reported record of 75 members makes sense. The visibility factor is low in such surveys. The investigator has surely missed most of them that are out there. Thus exploitation should not be a cause for alarm.

**ENVIRONMENTAL LOBBY:** The reported record of 75 makes sense. The investigator is an expert in such surveys. He has observed and recorded most that are there. And, therefore, only a few are out there. The population is threatened and needs to be protected.

The scenario is a typical one. It brings home the issues characteristic of field observations which often lack a sampling frame necessary for classical sampling theory to apply. One needs to work with visibility, catchability, audibility, etc. And, this is not a trivial problem!

2.3 Where Do They Go?

Someone concerned about the typical direction in which disoriented birds of a certain species fly goes out in an open field and stands facing north. He observes a bird vanish at the horizon at an angle of 10 degrees. A little later he finds a second bird vanish at the horizon at an angle of 350 degrees. He obtains the average \((10 + 350)/2 = 180\) degrees and declares that south is the typical direction. The diagram below, however, reveals that the above analysis is wrong. North is in fact a typical direction based on the evidence. What is wrong? And where? ... This extremely simple example has an astonishingly deep message in it: make sure that the modeling and analysis protocol does not mismatch the basics of the protocol of the phenomenon.
2.4 Martian Philosophy

A student wishes to study “Martian Philosophy” but finds that there is no instructional program available in Martian Philosophy. He is advised to take courses in astronomy, which may have some bearing upon Mars; he is also asked to take courses in philosophy that may have some context of the universe; and in due course, he is declared to have completed a program in Martian Philosophy! The inadequacy of this approach is clear. It would be important to make sure that neither the student nor the supervisor falls into this trap. Integrated and interactive research training programs should be made available to those interested and concerned.

2.5 Some Important Considerations

Our living resources are reasonably well described, and their value to our life support is reasonably well recognized. However, the need for conservation and management most often is lost in the shambles of adversarial proceedings. Science generally becomes part of the problem, and often enough the “patsy”. It seems far easier to use science to obfuscate rather than clarify. We scientists can take credit for studies which define the productivity and value of the resource but must also accept the blame for being less than helpful in the public decision-making process.

The problem is twofold. First, we need to do a better job of defining and understanding the ecological processes with which we are faced. Second, we need to describe these processes, and the state of the environment in relation to them, in a manner that supports effective public debate and management. Statistical ecology is a way to promote both aspects.

Most every major issue in science, technology, and society involves decision making based on components, such as basic concepts, representative data collection, and valid conclusions. Most every decision making component has some element of opaqueness to it. Conceptualization usually becomes a multidimensional issue with varied opinions on the significance of both the number and the kinds of the conceptual dimensions desired. Data collection and management involve both logical and logistical issues and problems involving the observer, the observed and the observing. Drawing valid and relevant conclusions is still another important matter.

In the light of the advice received on a major issue, the decision maker usually finds himself moving from the state of confusion to clearer confusion or from the state of clarity to confused clarity, or both! Upon assuming the Office of the President of the United States, Harry Truman called in a chief advisor. The advisor, with great skill and detail, briefed the President. On the one hand, he said, this and this, but on the other hand, he said, that and that. And that is how it went for the full session. After the advisor left, the President remarked to his aide that what he needed was an advisor who was one-handed!
The above story, while simple, has a lot in it. Even on the issue of single versus several models and solutions! Whether by a single individual or several! Left-handed, right-handed, and which do the advising! And why not both-handed in one! These issues are very important, and they are not simple.

A traditional approach has largely used the concept of pairwise interaction among the research scientist, statistical scientist, and the resource manager, but has let the interaction of the statistical scientist and the resource manager be at a minimum. Many of us have witnessed the limitation of this approach on the emergence of useful information. We feel that a triad approach of simultaneous working interaction among the research scientist, the statistical scientist, and the resource manager is a suitable protocol for useful information to emerge in the days ahead. The following schematic diagram encapsulates the idea.

![Schematic Diagram]

Useful Information for the Decision Maker

2.6 ORGANIZATIONAL HOMES AND FORUMS

Both statistical ecology and environmental statistics need, and richly deserve, appropriate homes and forums. We know that EPA has now a Statistical Policy Branch. It is time that the Ecological Society of America has a Statistical Ecology Section and that the American Statistical Association has an Environmental Statistics Section.

3. BIASED SAMPLING, WEIGHTED DISTRIBUTIONS APPROACH, AND UNBIASED INFERENCE

3.1 PROBING ENCOUNTERED DATA USING WEIGHTED DISTRIBUTIONS

Traditional statistical theory and practice have been occupied largely with statistics involving randomization and replication. But in ecological and environmental work, observations most often fall in the non-experimental, non-replicated, and non-random categories. Additionally, the problems of model specification and data interpretation acquire special importance and
great concern. In several situations of diverse nature in statistical ecology 
and environmental statistics, the theory of weighted distributions provides 
a perceptive and unifying approach for the problems of model specification 
and data interpretation.

Weighted distributions take into account the observer-observed inter-
face, i.e., the method of ascertainment, by adjusting the probabilities of 
actual occurrence of events to arrive at a specification of the probabilities of 
those events as observed and recorded. Failure to make such adjustments 
can lead to wrong conclusions.

The concept of weighted distributions can be traced to the study of the 
effect of methods of ascertainment upon estimation of frequencies by Fisher 
(1934). In extending the basic ideas of Fisher, Rao (1965) saw the need for 
a unifying concept and identified various sampling situations that can be 
modeled by what he called weighted distributions. Within the biomedical 
context of cell kinetics and the early detection of disease, Zelen (1974) in-
troduced weighted distributions to represent what he broadly perceived as 
length biased sampling. The initial idea of length biased sampling appears in 
Cox (1962).

Patil (1984) has discovered weighted distributions as stochastic models 
in the equilibrium study of populations subject to harvesting and predation. 
Mahfoud and Patil (1981) and Patil et al. (1986) have identified a Bayesian 
analogue to the theory of weighted distributions through the relationship of 
the posterior distribution to the prior distribution via the likelihood. Several 
papers have appeared that deal with weighted distributions and their appli-
cations in a variety of fields. For the growing literature, see the bibliography 

3.2 Univariate Weighted Distributions

Suppose $X$ is a non-negative observable random variable (rv) with its 
natural probability density function (pdf), $f(x; \theta)$, where the natural param-
eter $\theta$, is an element of the parameter space, $\Omega$. Suppose a realization $x$ of $X$ 
under $f(x; \theta)$ enters the investigator's record with probability proportional 
to $w(x, \beta)$, so that

$$
\frac{\text{Prob}(\text{Recording} | X = y)}{\text{Prob}(\text{Recording} | X = x)} = \frac{w(y, \beta)}{w(x, \beta)}.
$$

Here the recording (weight) function $w(x, \beta)$ is a non-negative function with 
parameter $\beta$ representing the recording (sighting) mechanism. Clearly, the 
recorded $x$ is not an observation on $X$, but on the rv $X^w$, say, having pdf,

$$
f^w(x; \theta, \beta) = \frac{w(x, \beta)f(x; \theta)}{\omega},
$$

(1)
where $\omega$ is the normalizing factor obtained to make the total probability equal to unity by choosing $\omega = E[w(X, \beta)]$. The rv $X^w$ is called the weighted version of $X$, and its distribution in relation to that of $X$ is called the weighted distribution with weight function $w$. Note that the weight function $w(x, \beta)$ need not lie between zero and one, and actually may exceed unity, as for example when $w(x, \beta) = x$, in which case, $X^* = X^w$ is called the size-biased version of $X$. The distribution of $X^*$ is called the size-biased distribution with pdf

$$f^*(x; \theta) = \frac{xf(x; \theta)}{\mu},$$

where $\mu = E[X]$. The pdf $f^*$ is called the length-biased or size-biased version of $f$, and the corresponding observational mechanism is called length or size biased sampling.

The concept of weighted distributions has been used during the last 25 years as a useful tool in the selection of appropriate models for observed data, specially when samples are drawn without a proper frame.

In many situations the model given in (1) is appropriate, and the statistical problems that arise are the determination of a suitable weight function $w(x, \beta)$ and the drawing of inferences on $\theta$.

The following examples from Patil and Rao (1977) may help illustrate a few situations generating weighted distributions.

I) **Truncation**: The distribution of a random variable truncated to a set $T$ is a weighted distribution with weight function $w(x) = 1$ for $x \in T$ and zero elsewhere.

II) **Damaged Observations**: Consider a damage model where an observation $X = x$ is reduced to $y$ by a destructive process with pdf $d(y|x)$. See Rao (1965). Then the probability that the observation $X = x$ is undamaged is $d(x|x)$, and the distribution of the undamaged observation is the weighted distribution with $w(x) = d(x|x)$. For example, under a binomial survival model, $d(x|x) = \theta^x$, $0 < \theta < 1$. An investigator recording only undamaged observations will need to work with a corresponding weighted distribution.

III) **Analysis of Family Data**: Various demographic and social studies involve family size and sex ratio as important factors which have some bearing on the main study. This example shows how a weighted distribution arises as a result of the size-biased sampling. The discussion is based on Rao (1965).

Consider the following data which relate to brothers and sisters in families of 104 boys who were admitted to a postgraduate course at the Indian Statistical Institute.
Let us assume that in families of given size $n$, the probability of a family with $x$ boys coming into the record is proportional to $x$. Also, suppose that the number of boys follows a binomial distribution with probability parameter $\pi$. Then

$$ f(x; \pi) = \binom{n}{x} \pi^x (1 - \pi)^{n-x} $$

$$ w(x) = x \quad , \quad E[w(X)] = \omega = n\pi $$

$$ f^w(x; \pi) = \binom{n-1}{x-1} \pi^{x-1} (1 - \pi)^{n-x} $$

$$ E[X^w/n] = \pi + (1 - \pi)/n > \pi $$

$$ E[(x^w - 1)/(n - 1)] = \pi \ . $$

If $k$ boys representing families of size $n_1, n_2, \ldots, n_k$ report $x_1, x_2, \ldots, x_k$ boys, an unbiased estimate of $\pi$ is

$$ \tilde{\pi} = \frac{\sum x_i - k}{\sum n_i - k} = \frac{414 - 104}{726 - 104} = \frac{1}{2} \ . $$

3.3 Weight Functions and Properties

Various forms of weight functions $w(x)$ have appeared in the scientific and statistical literature. Most of these weight functions are monotone increasing or decreasing functions of $x$. The following results provide useful comparisons of $X^w$ with $X$.

**Result 3.3.1** (Patil, Rao and Ratnaparkhi 1986): The weighted version $X^w$ is stochastically greater or smaller than the original rv $X$ according as the weight function $w(x)$ is monotone increasing or decreasing in $x$. As a consequence, the expected value of the weighted version $X^w$ is greater or smaller than the expected value of the original rv $X$ according as the weight function $w(x)$ is monotone increasing or decreasing in $x$.

**Result 3.3.2** (Zelen 1974; Patil and Rao 1978): The expected value of the size-biased version $X^*$ is $E[X^*] = \mu[1 + \sigma^2/\mu^2]$, where $E[X] = \mu$ and $V(X) = \sigma^2$. Further, the harmonic mean of $X^*$ is equal to the mean of the original rv $X$ when it is positive, i.e. $E[1/X^*] = 1/\mu$. Another way of expressing these results is: $E[X^*]E[1/X^*] = 1 + \sigma^2/\mu^2$. 
RESULT 3.3.3 (Patil and Ord 1975): Let the rv $X$ have pdf $f(x; \theta)$ and have size bias with weight function $x^{\beta}$. Then a necessary and sufficient condition for $f^w(x; \theta, \beta) = f(x; \eta)$, where $\eta = \eta(\theta, \beta)$, is that $f(x; \theta) = x^{\theta}a(x)/m(\theta)$. In this case, $f^w(x; \theta, \beta) = f(x; \theta + \beta)$. This result holds under certain mild regularity conditions.

RESULT 3.3.4 (Mahfoud and Patil 1981): Consider $X$ to be a rv $X$ subject to a size-bias with the weight function $x^\beta$. Then $X$ is log-normally distributed if and only if $V(\log X) \equiv V(\log X^w)$ for $\beta > 0$, where $V$ stands for variance. Thus the invariance of the logarithmic variance under size-bias of order $\beta$ characterizes the lognormal distribution.

EXAMPLE: In sedimentology and various other fields, the distribution of particles is usually analyzed by mass rather than frequency. Sieve analysis is a good example, which provides data consisting of sizes of sieves and corresponding masses of all particles retained by those sieves. It is interesting to note that the mass-size density is nothing but the weighted version of size-bias of order 3 of the pdf of the particle size. It can be verified that, if $X$ is lognormal with parameters $\mu$ and $\sigma^2$, then $X^w$ with $\beta = 3 \sigma^2$ is lognormal with parameters $\mu + 3 \sigma^2$ and $\sigma^2$. This property was empirically noticed and utilized for inference in the sedimentology literature. See, for example, Krumbein and Pettijohn (1938), and Herdan (1960).

3.4 POSTERIOR AND WEIGHTED DISTRIBUTIONS

There is a Bayesian analogue to the theory of weighted distributions. (See Mahfoud and Patil 1981; Patil, Rao and Ratnaparkhi 1986).

RESULT 3.4.1 (Mahfoud and Patil 1981): Consider the usual Bayesian inference in conjunction with $(X, \theta)$ having joint pdf $f(x, \theta) = f(x|\theta)f(\theta) = f(\theta|x)f(x)$. The posterior pdf

$$f(\theta|x) = f(x|\theta)f(\theta)/f(x) = \ell(\theta|x)f(\theta)/E[\ell(\theta|x)]$$

is a weighted version of the prior pdf $f(\theta)$. The weight function is the likelihood function of $\theta$ for the observed $x$.

RESULT 3.4.2 (Patil, Rao and Ratnaparkhi 1986): Consider the usual Bayesian inference in conjunction with $(X, \theta)$ with pdf $f(x, \theta) = f(x|\theta)f(\theta) = f(\theta|x)f(x)$. Let $w(x, \theta) = w(x)$ be the weight function for the distribution of $X|\theta$, so that the pdf of $X|\theta$ is $w(x)f(x|\theta)/w(\theta)$, where $w(\theta) = E[w(X)|\theta]$. Then the original and the weighted posteriors are related as follows:

$$f(\theta|x) = \frac{\omega(\theta)f^w(\theta|x)}{E[w(\theta)|X^w = x]}.$$
Further, the weighted posterior \( r v \theta^w|X^w = x \) is stochastically greater or smaller than the original posterior \( r v \theta|X = x \) according as \( \omega(\theta) \) is monotonically decreasing or increasing function of \( \theta \).

Some examples are given in tabular form as follows:

| \( X|\Theta = \theta \) | \( \Theta \) | \( \Theta|X = x \) | \( W(x) \) | \( [\Theta|X^w = x] \) |
|-------------------------|-------------|-----------------|--------|-----------------|
| Poisson(\( \theta \))   | Gamma(\( k, \lambda \)) | Gamma(\( k + x, \frac{1}{\lambda + 1} \)) | \( x \) | Gamma(\( k + x - 1, \frac{1}{\lambda + 1} \)) |
| Binomial(\( n, \theta \)) | Beta(\( a, b \)) | Beta(\( x + a, n - x + b \)) | \( x \) | Beta(\( x - 1 + a, n - x + b \)) |
| Neg-Bin(\( k, \theta \)) | Beta(\( a, b \)) | Beta(\( k + a, x + b \)) | \( x \) | Beta(\( k + 1 + a, x - a + b \)) |
| Exponential(\( \theta \)) | Gamma(\( k, \lambda \)) | Gamma(\( k + 1, \frac{1}{\lambda + 1} \)) | \( x \) | Gamma(\( k + 2, \frac{1}{\lambda + 1} \)) |

3.5 Applications of Weighted Distributions

A vast number of situations arise in which weighted distributions find application. These include: cell cycle analysis and pulse labeling, Zelen (1974); efficacy of early screening for disease and scheduling of examinations, Zelen (1971, 1974); cardiac transplantation, Temkin (1976); estimation of antigen frequencies, Simon (1980); ascertainment studies in genetics, Rao (1965,1985), Stene (1981); renewal theory and reliability, Cox (1962), Zelen (1974); non-renewable natural resource exploration, Barouch et al. (1985); traffic research, Brown (1972); word association analysis, Haight and Jones (1974); marketing and resource utilization, Morrison (1973); analysis of spatial pattern, Pielou (1977); species abundance and diversity, Engen (1978); transect sampling, Cook and Martin (1974), Gates (1979), Patil and Rao (1978), Patil, Taillee and Wigley (1979), Quinn (1979); forest products research, Warren (1975); income inequality and species inequitability, Hart (1975), Taille (1979); canonical hypothesis in ecology, Preston (1962), Patil and Taille (1979); particle size statistics, Gy (1982), Krumbein and Pettijohn (1938); mass-size distributions, Herdan (1960) and Schultz (1975); and quality of Swiss cheese, Tallis (1970).

4. Stochastic Population Dynamics, Weighted Distributions, and Recruitment in Fisheries

4.1 Weighted Distributions in Stochastic Population Dynamics

Consider the stochastic differential equation,

\[
\frac{1}{x} \frac{dx}{dt} = r(x, t) = g(x) + \gamma(t),
\]

where \( g(x) \) is the deterministic growth rate, \( \gamma(t) \) is the random perturbation, and \( x \) is the population size.
where,
\[ x(t) = \text{the population size at time } t, \]
\[ r(z, t) = \text{the per capita growth rate of a population of size } z \text{ at time } t, \]
\[ g(x) = \text{the biological part of the per capita growth rate dependent on population size } x, \text{ but independent of time } t, \text{ and} \]
\[ \gamma(t) = \text{the environmental part of the per capita growth rate dependent on time } t, \text{ but independent of the population size } z. \text{ Let } \gamma(t) \text{ be a white noise process with environmental unpredictability parameter } \sigma^2. \]

The population size \( x(t) \) is then a stochastic integral, and when it exists, its equilibrium pdf,
\[
f(x) = \frac{\text{const}}{V(x)} \exp \left[ 2 \int \frac{M(x)}{v(x)} \, dx \right] \\
= \exp[a \log x + b(x) + c],
\]
is a member of the log-exponential family, where \( M(x) = x g(x) \) and \( V(x) = \sigma^2 x^2 \).

Further, if the population is subjected to exploitation (harvesting, predation, etc.) with per capita exploitation rate \( h(x) \), the equilibrium pdf \( f_h(x) \) of the exploited population size interestingly simplifies to
\[
f_h(x) = \frac{w(x) f(x)}{E[W(x)]}
\]
where \( f(x) \) is the natural population equilibrium pdf and
\[
w(x) = \exp \left[ -2 \int \frac{x h(x)}{\sigma^2 x^2} \, dx \right].
\]

If the biological part of growth rate \( g(x) \) decreases linearly, hyperbolically or logarithmically, the equilibrium distributions turn out to be gamma, beta or lognormal respectively. Further, a constant harvesting rate \( h(x) \) gives a power function weight function, \( w(x) = x^\rho \). Also, if the harvesting rate \( h(x) \) is hyperbolic (Holling type II) given by \( h(x) = \lambda/(x + \alpha) \), then \( w(x) = [1 + (\alpha/x)]^\nu \), with \( \nu = 2\lambda/\alpha\sigma^2 \), and if it is sigmoid (Holling type III) given by \( h(x) = \lambda x/(\alpha^2 + x^2) \), then \( w(x) = [\exp \tan^{-1}(x/\alpha)]^\nu \), with \( \nu = -2\lambda/\alpha\sigma^2 \). For more discussion, see Patil (1984) and Dennis and Patil (1984).

4.2 Modeling and Analysis of Recruitment Distributions

Hennemuth, Palmer, and Brown (1980) reviewed recruitment data for eighteen fish stocks around the world by examining the observed frequency
distributions. The authors pointed out some indications of possible multi-modality in several of the data sets.

Taillie, Patil and Hennemuth (1988) continued the statistical analysis of the recruitment data of these eighteen fish stocks. The data were examined from several vantage points, such as skewness, kurtosis, and bimodality. Table 1 provides the distributions used with their density functions, ranges, and parameters. Distributions 3 and 4 have range \((-\infty, \infty)\) and are intended to represent the distribution of the logarithm of recruitment, \(\ln X\). The other eight distributions apply to \(X\) itself.

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
<th>Density Function</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>Lognormal</td>
<td>(\frac{1}{\sigma \sqrt{2\pi}} \exp -\frac{1}{2} \left( \frac{\ln x - \mu}{\sigma} \right)^2), (x &gt; 0)</td>
<td>(-\infty &lt; \mu &lt; \infty) (\sigma &gt; 0)</td>
</tr>
<tr>
<td>(2)</td>
<td>Gamma</td>
<td>(\frac{k^\lambda}{\Gamma(k)} x^{k-1} \exp(-\lambda x)), (x &gt; 0)</td>
<td>(k, \lambda &gt; 0)</td>
</tr>
<tr>
<td>(3)</td>
<td>Type (N) Catastrophe Model</td>
<td>Standardized pdf: (\exp(-A_0 + A_1 x + A_2 x^2 + A_3 x^3 + A_4 x^4))</td>
<td>(A_4 &gt; 0)</td>
</tr>
<tr>
<td>(4)</td>
<td>Mixture of Two Homoscedastic Normals</td>
<td>Standardized pdf: (\theta p \cdot n(x; \mu, \sigma, \delta, \lambda) + (1 - p) n(x; \mu, \sigma, \delta, \lambda))</td>
<td>(0 &lt; p &lt; 1)</td>
</tr>
<tr>
<td>(5)</td>
<td>Beta Type Two with a Scale Parameter</td>
<td>(\frac{1}{B(\alpha, \beta)} x^{\alpha-1} \exp(-\frac{\beta}{\alpha} x)), (x &gt; 0)</td>
<td>(\alpha, \beta, \lambda &gt; 0)</td>
</tr>
<tr>
<td>(6)</td>
<td>Symmetric Beta Type Two with a Scale Parameter</td>
<td>(\frac{1}{B(\alpha, \beta)} x^{\alpha-1} \exp(-\frac{\beta}{\alpha} x)), (x &gt; 0)</td>
<td>(\alpha, \lambda &gt; 0)</td>
</tr>
<tr>
<td>(7)</td>
<td>Reciprocal Gamma</td>
<td>* (\frac{1}{x \sigma} g\left(\frac{x}{\sigma}; k, \lambda\right))</td>
<td>(k, \lambda &gt; 0)</td>
</tr>
<tr>
<td>(8)</td>
<td>Weighted Gamma Type One</td>
<td>* (g(x; \lambda) \exp(-u_1(x))), (x &gt; 0)</td>
<td>(\beta, \lambda &gt; 0)</td>
</tr>
<tr>
<td>(9)</td>
<td>Weighted Gamma Type Two</td>
<td>* (g(x; \lambda) \exp(-u_2(x))), (x &gt; 0)</td>
<td>(\alpha, \lambda &gt; 0)</td>
</tr>
<tr>
<td>(10)</td>
<td>Weighted Gamma</td>
<td>* (g(x; \lambda) \exp(-u_3(x))), (x &gt; 0)</td>
<td>(k, \beta, \lambda &gt; 0)</td>
</tr>
</tbody>
</table>

\(\theta n(x; \mu, \lambda) = \sqrt{2\pi \lambda} \exp -\frac{1}{2}(x - \mu)^2 / \lambda; \) * \(g(x; \lambda)\) is the pdf of gamma in (2);

* \(u_1(x) = \frac{\sigma}{x} + \text{const.}; * u_2(x) = \frac{\beta}{x} + \frac{\sigma}{\lambda} + \text{const.}; * u_3(x) = \tan(\beta x) + \text{const.} \)

Table 1. Statistical distributions used for recruitment.
5. Modeling Clumped Sampling, Heterogeneity, and Extraneous Variation Using Weighted Distributions

5.1 Introduction

During their examination of the problem of toxoplasmosis, Diaconis and Efron (Diaconis and Efron 1985; Efron 1986) looked at the data sets involved and found there was more dispersion in the data sets than the existing models could accommodate and therefore they introduced a model called the double exponential family (DEF). This family, which is not to be confused with the Laplace distribution, enjoys the exponential family properties simultaneously for the mean and the dispersion parameters. It allows the data analyst to model overdispersion while carrying out the usual regression analyses for the mean as a function of the predictors. The overdispersion may be due to one or more possible causes, such as clumped sampling, heterogeneity, selection bias, etc. The issues pertaining to these kinds of causes arise in environmental and ecological research and management. Examples include soil sampling, water sampling, litter sampling, publication bias, newsworthy events and unusual reports in the news media, etc.

5.2 Double Exponential Family as a Model for Clumped Data

We write the linear exponential family with mean \( \mu \) and variance \( V(\mu) \) in the form

\[
f_\mu(y) = \exp \{ yA(\mu) - B(\mu) + D(y) \}
\]

where \( \mu A'(\mu) - B'(\mu) \equiv 0 \) and \( A'(\mu) = 1/V(\mu) \). The double exponential family (DEF) is then defined by

\[
f_{\mu,\theta}(y) = c(\mu, \theta) \theta^{\frac{1}{2}} f_\mu(y)^\theta f_y(y)^{1-\theta},
\]

where the constant \( c(\mu, \theta) \) is a normalizing constant and the motivation is to use the densities (2) as constituents of a regression analysis, in which the unknown parameters \( \mu \) and \( \theta \) are estimated from the data.

We (Patil 1987; Patil and Taillie 1988c) observe the DEF to be a weighted distribution of the following form:

\[
f_{\mu,\theta}(y) = \frac{w_{\mu,\theta}(y) f_\mu(y)}{E[w_{\mu,\theta}(Y)]}
\]

where the weight function is given by

\[
w_{\mu,\theta}(y) = \left\{ \frac{f_y(y)}{f_\mu(y)} \right\}^{1-\theta}.
\]

Furthermore, the weight function has the following interesting form

\[
w_{\mu,\theta}(y) = \exp\{(1 - \theta)I(y, \mu)\},
\]
where $I(y, \mu)$ is the Kullback-Leibler distance function between $y$ and $\mu$ for the density function $f$. Notice that the Kullback-Leibler distance increases with the distance from the mean $\mu$, thus allowing a more distant observation larger weight and accommodating extra dispersion in the data set when $1 - \theta > 0$.

Figure 1 provides a picturesque representation of the makeup of the density function of the DEF at $x$. Suppose we observe $x$ as the first observation under the natural pdf $f_\mu(\cdot)$. Suppose there is some clumping and the additional $x$ values occur under the pdf $f_\theta(\cdot)$ with mean $x$ and not with mean $\mu$. Then the pdf of the DEF at $x$ turns out to be the $\theta$-geometric mean of $f_\mu(x)$ and $f_\theta(x)$ where $\theta$ is a measure of clumping. Figures 2 and 3 plot Poisson and double Poisson probability functions with $\mu = 10.5$ and $\theta = .1, .5$. The increasing dispersion of the double Poisson with decreasing $\theta$ is evident in these pictures. For comparison, we have also plotted the negative binomial probability function having the same mean and variance as the double Poisson. Figures 4 and 5 plot the corresponding logarithmic weight functions. Also shown in these plots is the Pearson logarithmic weight function, $(1 - \theta)(x - \mu)^2$, although the latter is too explosive in the right tail to be validly applied to the Poisson distribution. It is encouraging to observe that the DEF keeps the mean value about the same and yet accommodates overdispersion or underdispersion of the data set.

Table 2 provides a brief excerpt of the toxoplasmosis data which prompted Diaconis and Efron to introduce the double exponential family to deal with the issues arising from the overdispersion of the data set. Efron (1986) attributes the overdispersion to cluster sampling, but we have argued that heterogeneity (city-to-city variation) is a more likely explanation and that it can also be incorporated into the DEF framework (Patil and Taillie 1988c).

6. **META-ANALYSIS INCORPORATING HETEROGENEITY AND PUBLICATION BIAS WITH WEIGHTED DISTRIBUTIONS**

Meta-analysis consists of quantitative methods for combining evidence from different studies about a particular issue. Applications of meta-analysis as a tool for investigating scientific questions and for guiding public policy decisions are diverse and include, for example, the analysis of the efficacy of psychotherapy (Smith, Glass and Miller 1980), the assessment of human lung cancer risks from various environmental emissions (DuMouchel and Harris 1983), and the U.S. Department of Education study on school desegregation and black achievement. See also, for example, Fisher (1932), Hedges and Olkin (1985), Laird, Patil and Taillie (1988), Light and Pillemer (1984), Rosenthal (1984), and Wolf (1986).
**PDF $f_{\mu,\nu}$ of Double Exponential Family**

Figure 1. Graphs of the pdf's of the double exponential family.

**Comparison of Probability Functions**

Figure 2. Comparison of probability functions ($\nu = 0.5$).
Figure 3. Comparison of probability functions ($\nu = 0.1$).

Figure 4. Weight functions relative to the Poisson ($\nu = 0.5$).
6.1 An Example of Meta-Analysis Due to Iyengar and Greenhouse

Iyengar and Greenhouse (1988) consider the following example given in Hedges and Olkin (1985, p. 303). The data come from ten studies comparing the effects of experimental open classroom education with traditional education on student creativity and are reproduced in Table 3. Following Hedges and Olkin, Iyengar and Greenhouse assume that all studies are estimating the same effect size, denoted by \( \theta \). For the \( i \)th study, the second column gives the sample size, \( N_i \), in each of the two samples. The third column gives the effect size estimate, \( \hat{\theta}_i \), where \( \hat{\theta}_i \) is the difference between the means for each education type divided by a pooled estimate of the standard deviation. The fourth and fifth columns give, respectively, the \( t \)-statistics, \( t_i \), and the corresponding df, \( q_i = 2N_i - 2 \) for each study.

Denote the density of a noncentral \( t \) distribution with noncentrality \( \eta \) and with \( q \) df by \( f(t; \eta, q) \). With no selection bias in the reporting of the studies in Table 1, the \( t_i \) have density \( f(t; (N_i/2)^{1/2} \hat{\theta}, q_i) \) under assumptions of independence and normality. Letting the weight function \( w(t) \) model the selection bias in reporting a result for which the value, \( t \), of the \( t \)-statistic is
<table>
<thead>
<tr>
<th>City No.</th>
<th>Predictor (Rainfall)</th>
<th>Response Prop. Pos.</th>
<th>Original Sample Size $n_j$</th>
<th>Effective Sample Size $n_j\theta_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1735</td>
<td>.500</td>
<td>4</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1973</td>
<td>.300</td>
<td>10</td>
<td>6.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1620</td>
<td>.278</td>
<td>18</td>
<td>13.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>1796</td>
<td>.532</td>
<td>77</td>
<td>15.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>2292</td>
<td>.622</td>
<td>37</td>
<td>29.1</td>
</tr>
</tbody>
</table>

- To assess the effect of rainfall on the proportions positive.
- Cities with large sample sizes $n_j$ have large influence.
- The subjects may have been accrued in clumps, so the statistician should be using smaller values of $n_j$, not necessarily proportionately smaller in each city.
- Inadequacy of binomial model for this overdispersion.

**Table 2.** Toxoplasmosis data showing the proportions of subjects testing positive ($Y_j$) and numbers tested ($n_j$) in 34 cities (Efron 1978, 1986).

Observed, the likelihood function for $\theta$ is given by

$$L(\theta, w) = \frac{\prod_{i=1}^{10} f(t_i; (N_i/2)^{\frac{1}{2}}\theta, q_i) w(t_i)}{\prod_{i=1}^{10} A((N_i/2)^{\frac{1}{2}}\theta, w, q_i)}, \quad (1)$$

where

$$A(\eta, w, q) = \int_{-\infty}^{\infty} f(t; \eta, q)w(t)\, dt \quad (2)$$

Four of the ten studies in Table 3 yield results that are significant at the 0.05 level, so an appropriate weight function for these data should be non-zero everywhere. To examine different possible selection schemes and to examine the effect of the choice of the weight function upon the inferences.


Table 3. Studies of effects of open vs. traditional education on creativity (Iyengar and Greenhouse 1988).

<table>
<thead>
<tr>
<th>i</th>
<th>$N_i$</th>
<th>$\hat{g}_i$</th>
<th>$t_i$</th>
<th>$q_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>90</td>
<td>-0.583</td>
<td>-3.91</td>
<td>178</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>0.535</td>
<td>2.39</td>
<td>78</td>
</tr>
<tr>
<td>3</td>
<td>36</td>
<td>0.779</td>
<td>3.31</td>
<td>70</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>1.052</td>
<td>3.33</td>
<td>38</td>
</tr>
<tr>
<td>5</td>
<td>22</td>
<td>0.563</td>
<td>1.87</td>
<td>42</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>0.308</td>
<td>0.69</td>
<td>18</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>0.081</td>
<td>0.18</td>
<td>18</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>0.598</td>
<td>1.34</td>
<td>18</td>
</tr>
<tr>
<td>9</td>
<td>39</td>
<td>-0.178</td>
<td>-0.79</td>
<td>76</td>
</tr>
<tr>
<td>10</td>
<td>50</td>
<td>-0.234</td>
<td>-1.17</td>
<td>98</td>
</tr>
</tbody>
</table>

about $\theta$, Iyengar and Greenhouse consider the following two parametric families of weight functions:

$$w_1(x; \beta; q) = \begin{cases} \frac{|x|^\beta}{t(q, 0.05)^\beta}, & \text{if } |x| \leq t(q, 0.05), \\ 1 & \text{otherwise} \end{cases}$$

and

$$w_2(x; \gamma, q) = \begin{cases} e^{-\gamma x}, & \text{if } |x| \leq t(q, 0.05) \\ 1, & \text{otherwise} \end{cases}$$

The parameters $\beta$ and $\gamma$ are nonnegative and $t(q, 0.05)$ is the 0.05 two-sided critical point of the central $T$-distribution with $q$ degrees of freedom.

Both families of weight functions imply that all studies showing statistically significant results will be reported. In addition, these families have the following features: when $\beta$ and $\gamma$ are zero, the weight functions indicate no selection bias; when $\beta$ and $\gamma$ are infinite, the weight functions become the Hedges-Olkin scheme in which only statistically significant results are reported. The two weight function families differ only for nonsignificant results; $w_1$ says that the reporting probability increases as the outcome approaches statistical significance, whereas $w_2$ says that the reporting probability is constant for all non-significant results.

6.2 Publication Bias Versus Heterogeneity in the Iyengar and Greenhouse Example

weight functions and, in particular, to let the weight function depend on the \( p \)-value instead of the test statistic; (ii) to examine the feasibility of using a normal approximation to the noncentral \( T \)-distribution; and (iii) to incorporate heterogeneity of the effects parameter, as well as publication bias, into the model. Heterogeneity and publication bias are two competing explanations for the overdispersion in the \( t \)-scores of Table 3, and a joint model would permit likelihood ratio tests of their relative importance. For this particular data set, we find that the normal approximation is quite adequate, and that heterogeneity accounts for nearly all of the overdispersion with virtually no evidence of publication bias. Further investigation with additional data sets would be needed to judge the general suitability of the normal approximation and to determine if such decisive discrimination between heterogeneity and publication bias is always possible. Sensitivity to the choice of weight function is largely an open question since we found no evidence of publication bias for these data.

**Weight Functions Examined:**

\[
 w_1 = \begin{cases} 
 (x/	ext{xcrit})^\alpha & \text{if } |x| < \text{xcrit} \\
 1 & \text{otherwise}
\end{cases}
\]

Half-Normal = \( \exp[-\beta^* p(x)^2] \)

Negative Exponential = \( \exp[-\beta^* p(x)] \)

Here \( p(x) \) is the \( p \)-value when the test statistic takes value \( x \). The weight function \( w_1 \) is the same as that considered by Iyengar and Greenhouse (cf. equation 3). The corresponding weighted distributions, as estimated from the data, are plotted in Figure 6 against the \( p \)-values and also against the value of the test statistic (see below for descriptions of the different models employed for parameter estimation). Asterisks along the horizontal axes indicate the ten observed \( p \)-values/\( T \)-scores.

### 6.3 Joint Weight Function for Heterogeneity and Publication Bias

In case of normal distributions, heterogeneity can be represented by a Pearson type of weight function

\[
 w_H(x) = e^{a(x-\theta)^2}, \quad a > 0, \quad (5)
\]

where \( \theta \) is the mean of the original normal distribution. As discussed in Section 5, Diaconis and Efron have incorporated overdispersion into more general (non-normal) exponential families by replacing \((x-\theta)^2\) in \((5)\) by the
Figure 6. Weighted distributions plotted against p-values and against test statistics for models 1-3.
Kullback-Liebler distance function. On the other hand, the various weight functions for publication bias can be written in a unified form as

\[ w(x) = w(x, \beta) = \exp(-\beta K(x)) \]

(6)

where

\[ K(x) = \log(x/\text{crit}) \quad \text{for } w = w_1 \]
\[ K(x) = p(x) \quad \text{for neg. exp.} \]
\[ K(x) = p(x)^2 \quad \text{for half-normal.} \]

Note that \( K(x) \) may have an implicit dependence on the degrees of freedom — either through the critical point, \( x_{\text{crit}} \), or through the \( p \)-value, \( p(x) \).

Multiplying (5) by \( \exp(-\beta K(x)) \), the combined effect of heterogeneity and publication bias is expressed by a joint weight function,

\[ w(x) = e^{\alpha(x-\theta)^2 - \beta K(x)} . \]

(7)

This form may account for our success in distinguishing heterogeneity from publication bias, at least when \( \theta \) is small. For large \( x \), \( \beta K(x) \approx 0 \) so the quadratic is the dominant term in (7). Conversely, \( \beta K(x) \) is the more important term for small \( x \). Thus, publication bias can be described as a short-range effect while heterogeneity is long-range.

7. Effects of Biased Sampling and Model Weight Functions on the Amount of Information

7.1 Introduction

In sections 2 and 3 we discussed situations in which the method of ascertainment gives rise to observations whose distribution is a weighted version of the original distribution. While it might be supposed that such "biased" sampling is disadvantageous for purposes of inference on the population parameters, this is not always the case. This section describes circumstances in which the sampling bias, when properly accounted for, leads to stronger inferences than would be possible with an equal number of observations from the original distribution. Patil and Taillie (1988b) consider observations from a one-parameter exponential family

\[ f(x) = \frac{c(x)e^{-\theta A(x)}}{M(\theta)} \]

(1)

and an equal number of observations from the \( w(x) \)-weighted version of this family

\[ f_w(x) = \frac{w(x)c(x)e^{-\theta A(x)}}{M_w(\theta)} . \]

(2)
The intent is to compare the relative information content, in the sense of Fisher, of the two sets of observations for inferences concerning \( \theta \). In particular, which of the two families is more informative for \( \theta \)? Also, for given \( w(x) \), when is it the case that \( f \) and \( f_w \) are equally informative for \( \theta \)? – a situation that may be described as Fisher neutral.

For family (1), the observed and expected Fisher's information per observation are equal and are given by

\[
J(\theta) = \frac{d^2}{d\theta^2} \log M(\theta).
\]

Thus the change in Fisher's information due to weighting becomes

\[
J_w(\theta) - J(\theta) = \frac{d^2}{d\theta^2} \log M_w(\theta) - \frac{d^2}{d\theta^2} \log M(\theta)
\]

\[
= \frac{d^2}{d\theta^2} \log \frac{M_w(\theta)}{M(\theta)}.
\]

Bayarri and DeGroot (1986) have obtained (4) for selection (i.e. truncation) models in which case \( M_w(\theta)/M(\theta) \) reduces to the selection probability. In general, \( M_w(\theta)/M(\theta) = E[w(X)] \) where the expectation is under model (1).

Equation (4) allows us to conclude that:

(a) The weighted version of \( f \) is uniformly more informative for \( \theta \) if and only if \( M_w(\theta)/M(\theta) \) is log convex.

(b) The weighted version of \( f \) is uniformly less informative for \( \theta \) if and only if \( M_w(\theta)/M(\theta) \) is log concave.

(c) \( f \) and \( f_w \) are uniformly equally informative (Fisher neutral) if and only if \( M_w(\theta)/M(\theta) \) is log-linear. For given \( w \), this characterizes Fisher neutrality by a functional equation involving \( M \).

Presumably there would be examples in which \( f \) is more informative for some \( \theta \) and less informative for other \( \theta \).

For the most part, we consider the size-biased weight function \( w(x) = x \) and assume that (1) is the density function for a non-negative random variable with respect to either Lebesgue measure or counting measure. Other weight functions do not introduce anything new in concept though they may be of practical interest. For example, the weight function \( w(0) = 0, w(x) = 1, x > 0 \), allows one to compare the information content of zero-truncated and untruncated observations from a discrete distribution. Bayarri and DeGroot (1986) give an extensive treatment of single-parameter truncation models from the perspective of Blackwell sufficiency as well as Fisher information.

From (1) we see that \( M(\theta) \) is a generalized transform of \( c(x) \) of a type determined by \( A(x) \). Thus when \( A(x) = x \) (linear exponential family), \( M(\theta) \)
is the Laplace transform of \( c(x) \), and when \( A(x) = \log(x) \) (log-exponential family), \( M(\theta) \) is the Mellin transform. The weighting of the distributions, \( f(x) \to f_w(x) \), corresponds to a certain operation, \( M(\theta) \to M(\theta) \), on the transforms. The change in information content attributable to weighting is a characteristic of this operation as measured by \( (d^2/d\theta^2) \log M_w(\theta)/M(\theta) \).

In general, the operation \( M(\theta) \to M(\theta) \) may be difficult or impossible to describe analytically. However, in the two cases considered here (linear exponential family and log-exponential family) the operations are differentiation and translation, respectively, when \( w(x) = x \).

### 7.2 Log Exponential Family

The log exponential family has pdf given by

\[
    f(x) = \frac{c(x)e^{\theta \log x}}{M(\theta)} = \frac{x^\theta \cdot c(x)}{M(\theta)}
\]

where \( M(\theta) \) = Mellin transform of \( c(x) \).

**Theorem:** Suppose \( X \) follows the log exponential family (5). Then for arbitrary weight function \( w \),

\[
    J_w(\theta) - J(\theta) = \text{var}(\log X_w) - \text{var}(\log X) \quad (6)
\]

In particular, the variable with larger logarithmic variance is more informative for \( \theta \).

For the remainder of this section we consider only the size-biased weight function, \( w(x) = x \), for which \( M_w(\theta) = M(\theta + 1) \).

### 7.3 Lognormal Distribution

\[
    f(x) = \frac{1}{\sqrt{2\pi \sigma^2}} \frac{1}{x} e^{\frac{1}{2\sigma^2}(\log x - \mu)^2} \quad (7)
\]

\[
    = \frac{1}{\sqrt{2\pi \sigma^2}} \frac{1}{x} e^{\frac{1}{2\sigma^2}(\log x)^2} \cdot \frac{e^{\frac{\mu^2}{2\sigma^2}} \log x}{e^{\frac{\mu^2}{2\sigma^2}}} \quad (8)
\]

When \( \sigma^2 \) is known, (8) is a log exponential family with \( \theta = \mu/\sigma^2 \) and

\[
    M(\theta) = e^{\frac{\theta^2}{2\sigma^2}} = e^{\frac{\mu^2}{2\sigma^2} \theta^2}.
\]

Thus

\[
    M_w(\theta) = M(\theta + 1) = e^{\frac{\theta^2}{2}(\theta + 1)^2}
\]

and
ENCOUNTERED DATA

\[ \frac{M_\omega(\theta)}{m(\theta)} = e^{\frac{\theta^2}{2}(2\theta+1)} \]

Since this is log linear in \( \theta \), it follows that the lognormal is Fisher-neutral for \( \mu/\sigma^2 \). This has a simple interpretation since the effect of the weight function is to translate the corresponding normal curve to the right by an amount equal to \( \sigma^2 \).

We conclude this section with a curious example in which size-biasing sometimes increases information and sometimes decreases information depending upon the value of a second parameter. Consider again the lognormal distribution (8) and let \( \theta = \mu/\sigma^2 \) and \( \phi = 1/\sigma^2 \). Regard \( \theta \) as known and \( \phi \) as unknown. The pdf is

\[ f(x) = \frac{1}{\sqrt{2\pi} x} \frac{1}{\sqrt{\phi} e^{\frac{\phi x^2}{2}}} e^{-\phi \frac{(\log x)^2}{2} + \phi \log x} \]

so

\[ M(\phi) = \frac{e^{\frac{\phi}{2\phi}}}{\sqrt{\phi}} \]

\[ M_\omega(\phi) = \frac{\exp\left\{ \frac{(\theta+1)^2}{2\phi} \right\}}{\sqrt{\phi}} \]

\[ \frac{M_\omega(\phi)}{m(\phi)} = e^{\frac{2\theta+1}{2\phi}} \]

and

\[ \frac{d^2}{d\phi^2} \log \frac{M_\omega(\phi)}{m(\phi)} = \frac{d^2}{d\phi^2} \left( \frac{2\theta + 1}{2\phi} \right) = \frac{2\theta + 1}{\phi^3} \]

Positivity is equivalent to \( 2\theta + 1 > 0 \) which is equivalent to \( (\mu/\sigma^2) > -1/2 \). Note that Fisher neutrality occurs when \( (\mu/\sigma^2) = -1/2 \) which happens when the normal curves corresponding to \( X \) and \( X_\omega \) are symmetrically located with respect to the origin (Figure 7).

8. MAKING OF A CRYSTAL CUBE FOR COASTAL AND ESTUARINE DEGRADATION

8.1 INTRODUCTION

Environmental regulators and decision-makers would like to have a crystal ball that could predict how ecosystems would respond to stresses such as pollution or over-fishing. In this way, information on important parameters
such as effects on valued species, propagation of these effects through the ecosystem, and subsequent recovery after the removal of these stresses, could all be properly considered for the protection of important natural resources. In the real world, however, such information cannot be gained with certainty.

Instead a conceptual crystal cube having a series of faces, each of which represents a specific parameter that can be directly related to marine environmental degradation, has been constructed. At present, 10 indices or faces of the cube are being tested: dietary risks from contaminants in marine foods; contaminant stress in sediments; contaminant stress in the water column; human pathogen risks; benthic species distribution and composition; fish and shellfish diseases; reproduction in fish and shellfish; mortality in eggs and larvae of fish and shellfish; reproductive success in marine birds; and oxygen depletion effects (O'Connor and Dewling 1986).

The emphasis in developing and testing these indices is on standardizing long-term data sets in order to develop a single variable, termed an index. This index is based on a variable that measures contamination or, ideally, contamination effects. The choice of such a variable is not easy and usually involves extensive data analysis. To be useful, the index must also be sensitive to contamination and relatively insensitive to other factors. Selection of useful indices also requires consultation with their principal users, decision-makers and the public.

This variable index in the crystal cube analyses is used to separate "concern" or "alarm" from "no concern" conditions. Here, concern or alarm
indicates not only that legislated or regulated standards have been violated, but that the scientific community is unable to assure decision-makers that issues of widespread public concern will not arise from specific environmental stress.

The index is calibrated so that when the number falls in the range of 0 to 1, there is "no cause for concern." A flag is raised as soon as the index becomes 1. The range from 1 to 10 indicates "warning;" something is happening and should be investigated. The range above 10 indicates "cause for alarm." The index is designed to be 10 when there is scientific reason for ecological concern; the environment has been adversely affected.

The fundamental concept underlying the use of an indicator variable of environmental degradation is to compare conditions in a stressed estuary or coastal area with those of a clean region. The crystal cube with 10 faces, each representing an indicator variable of one important environmental pollution parameter, will flag cases where legal or scientific benchmarks are exceeded.

This technique will assist the environmental decision-maker, to focus attention on those specific environmental parameters indicative of serious problems. The crystal cube is not intended to be the "ideal" crystal ball desired by environmental managers, but it should provide a framework for evaluating and comparing different environmental measures that must be weighed not only against each other, but also against other (e.g., economic, aesthetic, etc.) considerations. Thus, the crystal cube could develop into a valuable tool to help define or delineate "unreasonable degradation" and make environmental decision-making more consistent and equitable.

Choices of the demarcation points between "no cause for concern" and "warning" and between "warning" and "alarm," at least ideally, require agonizing consensus among interested parties, the decision-maker, and scientists.

How do we choose the errors of the first and second kind $\alpha$ and $\beta$? We choose $\alpha$ to be 1 in 10 partly because ecological variability is rather large and also because a typical two-term ten year manager may be able to stand a false alarm once in a ten year period, which is also roughly a half human generation time. We choose detection power $1 - \beta$ to be 2 in 3 so that one is not caught napping in two successive years! For details, see O'Connor and Dewling (1986), O'Connor and Flemer (1986), Boswell and Patil (1985, 1986), O'Connor, Murchelano and Ziskowski (1987). In this section, we attempt a unified view of the ideas that have contributed to the making of the ten marine indices. The emerging insight and underlying unity are expected to be useful in a much wider context of index making and environmental management. For details, see Patil and Taillie (1987).
8.2 Indicator Variable, Its Benchmark, and the Index

A measure of the pollutant effect on which an index is based is called an indicator variable. A measure of reproductive success in marine birds would be an example.

Broadly speaking, an indicator variable is usually a suitable non-negative univariate summarization of a body of monitoring data. To be useful, it needs to have the following desirable properties:

**Property 1 (Associatedness):** It is associated with pollution, but is not necessarily a direct measurement of pollution itself.

**Property 2 (Directionality):** It increases with pollution. If a certain variable decreases with pollution, its reciprocal satisfies the directionality property.

**Property 3 (Sensitivity):** It is sensitive to the causative pollution factor, so that the indicator variable has the capacity to detect pollution. Further, it is insensitive to extraneous non-pollution, so that the indicator variable generates only limited "false alarms."

Sometimes a question is raised as to the need for indices when we have good measures of pollutant effects. Why not just interpret the measurements when making decisions? The answer lies in the realization that useful interpretation of the measurement value does not usually come about until one has begun to see the measurement value as a relative value — relative to some reference value. For example, as O'Connor (1984) has put it, "...The estimated prevalence of gastroenteritis from bathing has significance only relative to something...to guidelines or standards, to the costs of reducing risks for gastroenteritis, to the social benefits from swimming despite its disease risks, etc...Similarly, the practical significance of particular toxic concentrations in marine food species is not clear until we compare them with something — with governmentally determined acceptable concentrations, with revised assessments of dose-response in humans, etc..."

We would now wish to calibrate the indicator variable relative to some standard. But what yardstick do we choose and on what basis, so that the resultant indicator variable when measured in the units of the yardstick provides a useful index. O'Connor and Dewling (1986) propose that the yardstick be the benchmark or the reference value of the indicator variable that separates concern from no concern. The corresponding index then becomes

\[
\text{Index} = \frac{\text{Indicator Variable}}{\text{Benchmark}}.
\]
Note that the benchmark is nothing but a critical value in some sense. Also note that regardless of the nature of the indicator variable and regardless of the actual choice of its benchmark, the index so defined has a most desirable common feature in that this definition of the index calibrates the index in terms of degradation with index value unity (I = 1) separating concern from no concern.

Finally, how do we define and choose a benchmark?

8.3 Statistical Approach to Defining B

One of the required properties of the index is that it should not trigger false concern. In other words, the indicator variable should not be overly responsive to natural environmental factors that are unrelated to pollution. In reality, of course, any indicator variable will exhibit some sensitivity to non-pollutant factors. This sensitivity has to be "factored" out of the index.

Consider the distribution, $F_0$, of the indicator variable in an unstressed environment (Figure 8). The distribution $F_0$ describes the variability in $X$ due to natural environmental fluctuation. (In particular, $F_0$ is not a sampling distribution.) By contrast, in a stressed environment, the variable $X$ should have a distribution, $F_1$, that is shifted to the right. Here the meaning of "shifted" is left vague; it is not necessarily represented by a simple change of location or of scale. In any case, the greater the "degradation" the greater the shift in the region described by $F_1$.

![Figure 8. Distribution of indicator variable (measure of pollutant effect).]
Comparison of the preceding figures might suggest that an index could be developed on the basis of a hypothesis test with \( H_0 : F_1 = F_0 \) as the null hypothesis. However, hypothesis testing is concerned with proving (statistically) that a difference exists. The index needs to quantify that difference. Thus, the significance of a large index value is not so much that there is strong statistical evidence that \( F_0 \) differs from \( F_1 \). Instead, it means that pollutant stress is approaching a level of practical concern. The issue here is the distinction between statistical significance and practical importance.

Given that the indicator variable exhibits variability as described by \( F_0 \) under unstressed circumstances, how can we establish a benchmark by which the index can be calibrated? Clearly, a typical value such as the mean, \( \mu_{F_0} \), of \( F_0 \) is too small. What is needed is some plausible upper bound for \( X \) when pollution is not present. Several possibilities come to mind.

1. A suitable multiple, say \( c\mu_{F_0} \), of the mean. This would still leave us with the problem of determining \( c \). Clearly \( c \) should be large when \( F_0 \) has a large variance and, conversely, small when \( F_0 \) has a small variance.

![Figure 9. 90th percentiles of various distributions in terms of the coefficients of variation.](image)

2. A percentile of \( F_0 \). This relates directly to the rate of false concern. For example if the benchmark is taken as the 90th percentile of \( F_0 \), then realizations of \( X \) from \( F_0 \) will exceed the benchmark one time in ten — false concern 10% of the time.
If one compares (2) with the approach in (1), one may ask if there is a robust choice of $c$ so that $c \mu_{F_0} = 90$th percentile of $F_0$. Figure 9 shows the ratio $(90$th percentile$)/\mu_{F_0}$ plotted as a function of the coefficient of variation for five families of distributions. For the normal family, the ratio

$$\frac{90 \text{th percentile}}{\mu} = 1 + z_{10}, \text{CV}$$

is a linear function of the coefficient of variation whose slope is the standardized percentile $z_{10} = 1.282$. The other four families represent a variety of tail weights, ranging from light-tailed (gamma) to heavy-tailed (reciprocal gamma) with the lognormal intermediate. The SBeta distribution is the beta distribution of the second kind with equal index parameters and is of interest in this context because the logarithm is a symmetrizing transformation for this family. For the definition of these and other statistical distributions and for some of their important properties, see Table 1 and Patil et al. (1984).

Figure 9 exhibits some salient features that will be of value for estimation purposes:

a) The normal theory formula

$$90 \text{th percentile} = \mu + \sigma z_{10}$$

provides an excellent and robust approximation when the CV is not too large, say $CV < 60\%$. For light-tailed distributions, the approximation is satisfactory over a wider range ($CV < 150\%$).

b) For heavy-tailed distributions,

$$90 \text{th percentile} = 2 \cdot \mu_{F_0}$$

provides a reasonable approximation when CV is moderately large ($100\% < CV < 500\%$).

c) The ratio $c$ is decreasing for large CV. In effect, the mean "chases and eventually catches up" with the 90th percentile as the tail gets longer. Note, however, that the coefficient of variation must equal 2700% if the mean is to equal the 90th percentile for the lognormal.

(3) To the man in the street, a 10% rate of false concern means 1 in 10 which suggests that the benchmark might be defined in terms of order statistics. Perhaps as the largest of nine (conceptual) observations. A largest order statistic is a random variable, though, and the benchmark ought to be a parameter of the distribution, $F_0$. To obtain a parameter, one might consider the expected largest order statistic, i.e. the average in all possible conceptual draws of nine. Such a benchmark would be
very sensitive to the right hand tail of $F_0$. A long tail would result in a
large value for the benchmark. As a consequence, the rate at which false
concern would be indicated by the index would depend upon the shape
of $F_0$. Tabulated below are the false concern rates (as a percentage) for
several distributions and several coefficients of variation when $B$ equals
the expected largest of nine observations.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>10%</th>
<th>25%</th>
<th>50%</th>
<th>100%</th>
<th>150%</th>
<th>200%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>6.8</td>
<td>6.8</td>
<td>6.8</td>
<td>6.8</td>
<td>6.8</td>
<td>6.8</td>
</tr>
<tr>
<td>Gamma</td>
<td>6.7</td>
<td>6.5</td>
<td>6.3</td>
<td>5.9</td>
<td>5.6</td>
<td>5.3</td>
</tr>
<tr>
<td>Lognormal</td>
<td>6.6</td>
<td>6.3</td>
<td>5.8</td>
<td>5.0</td>
<td>4.5</td>
<td>4.1</td>
</tr>
<tr>
<td>Reciprocal Gamma</td>
<td>6.6</td>
<td>6.1</td>
<td>5.3</td>
<td>4.4</td>
<td>4.0</td>
<td>3.8</td>
</tr>
</tbody>
</table>

On the whole, defining the benchmark as a percentile seems the most
suitable approach since it allows one to control explicitly the rate of false
concern.

8.4 Estimation

We need to estimate the 90th percentile, $B$, of the reference distribution
$F_0$. We assume that a random sample $X_1, X_2, \ldots, X_n$ is available for this
purpose. In practice $n$ is usually small so we concentrate upon small sample
estimation.

As was pointed out earlier, when the coefficient of variation of $F_0$ is
not too large, the 90th percentile can be well approximated by the normal
testory expression.

\[ B = \mu = z\sigma \]  

where $z = z_{10} = 1.282$. An estimator of $B$ is then obtained by substituting
the sample mean and standard deviation for $\mu$ and $\sigma$:

\[ \hat{B} = \bar{X} + zS \]  

A question arises in connection with (2): should one use the normal $z$-
score or the t-score with $n-1$ degrees of freedom? The answer depends upon
where one's emphasis lies:

(A) Should $\hat{B}$ be close to $B$?
   i.e. $E[\hat{B}] = B$ (or possibly $E[1/\hat{B}] = 1/B$).

(B) Should the probability or false alarm be close to the targeted 10%? i.e.
    $E[F_0(\hat{B})] = 90\%$.  

If the emphasis is as in (A) then (assuming near normality) it is appropriate to use the z-score instead of the t-score. Indeed

\[ E[\hat{B}] = E[\bar{X} + ZS] = E[\bar{X}] + zE[S] \approx \mu + z\sigma. \]

Note that the last step is only an approximation since \( S \) slightly underestimates \( \sigma \) (see Fisher (1920) for a discussion of the bias correction).

On the other hand, a t-score is appropriate if the goal is to be close to the targeted false alarm rate. To see this, let \( X \) be a (hypothetical) new observation from \( F_0 \). Now we want

\[
\begin{align*}
E[F_0(\hat{B})] &= .90 \\
E[P(X < z|\hat{B} = z)] &= .90 \\
P(X < \hat{B}) &= .90.
\end{align*}
\]

(5)

Now consider estimates of the form \( \hat{B} = \bar{X} + kS \) where \( k \) is to be chosen so that (5) is satisfied. Then

\[ P(X < \bar{X} + kS) = .90 \]

or

\[ P \left( \frac{X - \bar{X}}{S} < k \right) = .90 \]

But

\[ \frac{1}{\sqrt{1 + \frac{1}{n}}} \frac{X - \bar{X}}{S} \sim T_{n-1} \]

so we may select

\[ k = T_{.10} \cdot \sqrt{1 + \frac{1}{n}}. \]

It may be noted that the estimate

\[ \hat{B} = \bar{X} + T_{.10} \cdot S \sqrt{1 + \frac{1}{n}} \]

is the upper 90 percent prediction limit for a new observation \( X \) (assuming normality), see Whitmore (1986).
Acknowledgments

Work on this paper has been carried out with partial support of research grants EPA CR-815273-01 and NOAA NA87-EAH-00009 at the Penn State Center for Statistical Ecology and Environmental Statistics, and NIH CA23415 at the Harvard Department of Biostatistics and the Dana Farber Cancer Institute. Harvard Biostatistics Working Group Sessions on Weighted Distributions and Meta Analysis have been particularly helpful. Also helpful were the discussions at the Mini-seminars of the Penn State Center for Statistical Ecology and Environmental Statistics.

References


O'Connor, J.S., R.A. Murchelano and J.J. Ziskowski (1987), "Index of Pollutant-Induced Fish and Shellfish Disease". NOAA Special Report, OMA.


**DISCUSSION**

_Joel O'Connor^2_

It is certainly a pleasure to discuss the paper by Professor Patil (GP) because the directions he has outlined already have borne fruit in several practical arenas and have a great deal of promise for both ecological research and environmental decision making. I want to emphasize this utility of the approaches that he has summarized. I should state at the outset that my perspective is that of an ecologist and of a science manager and distinctly not that of a statistician.

The crystal cube that GP mentioned is one example of a useful tool for environmental decision-making. The problem of unreasonable degradation that led to the crystal cube was raised by congressional committees in their frustration to interpret and make explicit several environmental laws passed during the 60s and early 70s. The point is made in several of these laws and the regulations that flow from them that thou shalt not unreasonably degrade the marine environment, the air or the water and the land. The frustration has always been how to quantify precisely what we mean by "unreasonable degradation." So, the crystal cube is an effort to attack this increasingly serious question, and I think it deserves, as GP says, hard looks

---

^2 Formerly a senior ecologist with the Ocean Assessments Division of the National Oceanic and Atmospheric Administration, is, at present, Ocean Policy Coordinator, EPA Region II, New York.
at existing data and perhaps even revision of conventional decision-making approaches. The first step that we took in a pretty large interdisciplinary context is a step often not taken in this sort of effort. We asked what marine impacts are most important in a social sense, not in a scientific or technical sense but in a social sense. We came up with a reasonable consensus that seems to be repeated commonly among the scientific community, decision-makers, environmentalists, and so forth.

The next step was to develop initial draft indices characterizing these particular environmental impacts. These were developed by marine scientists, lawyers, economists, environmentalists and decision-makers. Then we went to GP and his colleagues at the Penn State Center for Statistical Ecology and Environmental Statistics, initially in a fairly naive effort to get them to refine these a little bit perhaps and get their statistical certification that they were all right. Fortunately, they wouldn't have any of this statistical certification business, and took such a hard look at the initial drafts of the indices that we had developed, that we completely revised them and restructured them around the unifying principle that made all of the indices comparable. This restructuring made the approach to indexing a great deal more powerful and appealing to both the decision-making community and to the public. I think this is an excellent example of how statisticians can contribute to both science and management well beyond mere analysis of data. Statisticians can contribute to improve the fundamental approach to a problem that hasn't been seen by the scientists and the managers.

Almost any characterization of ecosystem structure requires encountered data. This is because there is seldom enough time to get very much new data before decisions need to be made. I am convinced that better interpretation of encountered data will improve environmental decisions perhaps as much as any other technical advance. The reasons for this are perhaps most importantly because ecosystems are so difficult to characterize. There are such great uncertainties in ecosystem behavior and so much structural variability in space and time. The indications are that these uncertainties will remain no matter how intensively we assess natural ecosystems.

GP has also emphasized size biased sampling of environmental variables. I think this is a typical problem with the variables used in the crystal cube indices. As an example, the fish disease index is based on the prevalence of different diseases of fishes in nominally contaminated and uncontaminated areas. We know that with essentially every species of fish individuals become diseased around urban areas and then migrate quite freely to uncontaminated areas. So the migrations of diseased fishes increase our apparent estimate of the disease prevalence way off the continental shelf someplace or in another uncontaminated area. This sort of example is very common.

Size-bias sampling may be typical for most ecosystem variables that are truly useful for environmental decision-making. For instance, the distribution and abundance of most organisms. Measures of most of these variables are inherently biased and correct interpretations require careful thought to
remove these biases. There are some indications that the directions and approaches outlined by GP are becoming recognized and applied to social issues. The American Fisheries Society has recognized GP’s work with Brad Brown of the National Marine Fisheries Service on risk analysis and the haddock fishery, and has given them a most significant paper award — the first time a mathematical statistician has received any such award of the American Fisheries Society.

Unfortunately, however, only a few ecologists or environmental decision makers are aware of these developments and their potential. I would also therefore emphasize the need for interaction and mutual education. It is time that both the Ecological Society of America and the American Statistical Association provide such a forum through a new Section as put forward by GP.

DISCUSSION

N. Phillip Ross

It is a pleasure to have this opportunity to discuss Dr. Patil’s work on encountered data. One of the problems that we face as statisticians involved in environmental decision making is that we’re often bearers of bad tidings. We’re often asked to come in after the information has been collected and the analysis is not quite accomplished, and asked to make some sense of it. And, what happens? (Or at least used to happen, as things are now changing partly due to some of the work that GP has been doing.) We’d look at this material and say there is nothing we can do... This is terrible... Throw it out and do it over again. When you tell a manager to throw out two million dollars worth of data and do it over again when his merit pay is dependent upon cost containment and analysis that reaches decisions in a timely manner, he throws you out!

As GP says, we are looking at a problem of soft data and we have to have some kind of a hard look at it. And, in the environmental area, most of the problems, at least in a regulatory context, that come up need immediate decisions. Also, we can’t design studies, in which we go out into the field and spend a year or two or three designing, collecting and analyzing information, and then give the decision to the administrator. A good example is Love Canal which EPA and other federal agencies got involved in. This is toxic waste. It was being dumped for years. People knew about it. They covered it up with dirt. Nobody set up any monitoring systems to see what

---

3 Chief of the Statistical Policy Branch in the Office of Policy, Planning and Evaluation of the United States Environmental Protection Agency, Washington, DC.
was happening to the waste material in terms of ground water, exposure for human health, etc. We didn't do that. It was only when material started oozing up out of the ground and people could smell it and see it in their basements and the newspapers got a hold of it that it became a very critical problem. This isn't to say that the scientists in the environmental agencies had their heads in the sand. They knew it was a problem. There were attempts to get monitoring systems set up and some types of controls set up. But, it really took this sort of catastrophic situation to raise the minds of Congress. In an environmental agency such as EPA, we only operate under statutes and directives from the Congress. As some of you might well be aware, when this problem occurred, we were given some very definite mandates to do something about it in a period of four months. We had to collect several thousand data points, analyze that data, go to the literature and look at the combined information for possible health effects of the toxins we were finding in the soil and in the air.

And, there have been two or three other Love Canal type of situations and we have learned a little bit from them. But, they all involve the use of looking at data that has already been collected. We don't have the time to collect good data. We look at what is happening in the area of human health through the literature. We don't have the chance to do human studies or animal studies when we have to make these decisions so we must look at what is available. We must look at very soft data, and develop very hard ways of looking at these things. I think that it's fairly apparent just from a Love Canal type story that environmental problems are extremely complex to assess, and there are not very many tools in the statistical arsenal to deal with some of the problems that we come in contact with, and as I indicated earlier, we are not in a position in many instances to design studies to collect data to use classical techniques for the analysis, to make determinations as to whether changes have or have not occurred.

Another example is the problem of hazardous waste sites in ground water monitoring. One of the things that EPA is charged with is to determine whether a hazardous waste site is contaminated or not. We do set up monitoring devices to detect whether or not there has been a breach in the lining or there is some kind of leaking into the aquifer of a hazardous waste site but these are not designed from a statistical point of view. They are designed really by convenience and from hydrological points of view; the statisticians in the agency are supposed to try to make sense out of this monitoring data and to tell management whether or not, there has been a breach or a problem at this site, or to give management an analytical tool that will tell them whether or not there has been a breach or a problem at this site. And, if there is, of course, this can cost the owner of the site his license if he can't clean it up, and it can cost millions of dollars to correct the problems. So, industry is quite concerned about the decision tools we use to impose on them whether or not they should clean up. And, we are quite concerned about the adequacy of the techniques like t-tests which have been proposed
in the public record. We, as the statistical component of the agency, are not comfortable with that at all.

Groundwater poses a very complicated problem. It is not a univariate problem although the tests that have been proposed are univariate. It is multivariate. And, most likely, multivariate non-parametric and we don't have a method. We don't have a way of dealing with this. And, we are never going to collect the kinds of data that we should if we could design it. It's an impossibility. We are left out of that process because of the realities of costs and the politics of doing things. So, we have to develop tools that will allow us to assess what we have even if it isn't very good data. And, we must be able to approach our management and work with them as GP pointed out. One of the most successful parts of the thinking that is going on in the encountered data area that is being picked up at EPA is forcing management to sit down with scientists and statisticians and talk about the problems. What are their objectives? How much power do they want in a test? Two out of three? or three out of three? These types of questions are being brought up and discussed at roundtables and it is beginning to work, and hopefully work well, for quick decision making that we expect to have to make in the future.

The other thing I would like to mention in closing is this need for new tools that we have recognized. I have also read the material that GP has given me in the past. Most of it looks to me interesting, relevant and on the cutting edge of statistical ecology and environmental statistics. I don't really understand it all yet. And, I'm sure that's not true for me alone. It is quite true for the managers that we try to sell it to. We try to have them agree to use it. This is not their field. They really don't understand it. They have to take it at face value from us. There has to be a way that we can convince them it works and show them it works better than what they are using now and improves upon the kinds of decisions they are making now. This emphasizes a need for the profession, the statistical profession, to recognize that environmental statistics is a discipline, an area which really needs to be recognized as a sub-discipline. One which requires new methods and tools to be developed. I agree with GP's position and would like to emphasize the need for ASA in particular to start looking at it as a separate section, and to start educating statisticians which in turn will help improve the kinds of environmental decisions we can make in concert with the management.
DISCUSSION

Morris H. DeGroot

I would like to congratulate GP on an excellent paper. In my brief comments, I would like to discuss some of the particular statistics problems that have interested Suzie Bayarri and myself over the last year or two. One very simple kind of weighted distribution is what we call a selection model and that's the particular problem in which the weight function is just 0 or 1. Then what we see is a random sample from this selection model and the problem would be to make appropriate inferences. One very simple but interesting problem is where we have a uniform distribution over some unknown set. We get a random sample so uniformly distributed you can picture it in the plane, for example, coming from some region with an unknown boundary. And, the problem is by looking at the points uniformly distributed to determine the boundary of that set from which they came. And in two dimensions or higher, it is not an obvious solution.

The second kind of example, in a medical context, is one in which we are interested in the distribution of some characteristic, $x$, over the general population but we can only observe the characteristic, $x$, in patients who exhibit a certain disease and the disease is defined by the property that the value of $x$ for those patients belongs to a specified set. Think of $x$ as some vector of enzymes, for example, or other biochemical variables and if that vector lies out in some extreme region then a person comes down with a certain disorder and it may be painful or expensive or dangerous to measure $x$ in a person. It may be only possible through some sort of surgical procedure and so we can only observe $x$ in patients on the basis of observing those patients, we want to make inferences about $x$ over the entire population. And, indeed, we may not know what the set $s$ is and in order to understand the disease better part of the problem may be in determining the region that causes this disorder to occur.

The statistics literature has many papers on inference about binomial or Poisson distributions with the zero class missing. That's the problem in selection models.

Perfect predictions? There is the well known old story that I unblushingly repeat to you now about the confidence racket where today I send out a couple of hundred letters to people out in the world and in half of the letters I say, 'You don't know me but I am an expert on the stock market and I have the ability to predict how the market is going to behave and out of the kindness of my heart I am sending you this free tip today. A certain stock, $A$, will rise in value during the coming week.' To the other half of

---

4 Now deceased, was University Professor of Statistics and Industrial Administration at Carnegie Mellon University, Pittsburgh, PA.
the people I send a letter that says almost the same thing but it happens to say that certain stock, $A$, will not rise in value during the coming week. Then I wait and see what happens during the week. To those half to whom I sent the correct prediction, I write again next Monday. And, I say, 'You may remember last Monday I wrote to you blah, blah, blah... and made a prediction and you may have noticed it came true and out of the kindness of my heart I am giving you another free prediction this week. Stock $B$ will go up in value.' I send that to half and to the other half I say 'Stock $B$ will not go up in value' and I just continue this way, week after week. After seven or so weeks I still have maybe ten people on my list to whom I have sent seven perfect predictions over those weeks and that's all they have seen, of course. Every Monday morning they get in their mail this mysterious letter making a correct prediction and I say, 'Well, I think I have proven to you that I can predict the stock market quite accurately and for $10,000 I will give you my next prediction.' Well, this is certainly an example of a selection model. And, I think, we all experience this often without realizing it.

In our more immediate daily lives, we may consider the news on TV or newspapers, a topic that GP has touched. Someone selects what they are going to show us in three minutes of news and twenty-seven minutes of commercials that we see each evening on TV and then our job from seeing that selection of the news is to get our picture of what's going on in the world. And, of course, as GP says only unusual events or extraordinary events get selected and only some of those and then we try to get a picture of how to try to proceed in our daily lives. This is, I think, an example that can be analyzed only from the Bayesian perspective. That is, we can only understand what the news means in our daily life when we have some prior experience about what is the ordinary and what is ongoing as a background on which to place this new selected data.

One very statistical example that is dear to our hearts is the following: The analysis of reports of significance levels in the scientific literature. We all are familiar with this basic statistical methodology.

Now, experimental results unfortunately get reported in the literature if they are significant much more easily than if they report nonsignificant results. So, what is the situation from the point of view of the reader of one of these journals? We sit down with our journal of science and read about the experiment that produced such an outcome. We would not be reading about this experiment at all unless it had produced a statistically significant outcome. So, from our point of view as a reader, the data have been selected. We have some interesting calculations about how you change your opinions about the natural parameters of interest in the light of this selection.

Let me put in a word about the Bayesian point of view on these problems. Here is the general weighted distribution model that GP has. The underlying distribution is $f(x; \theta)$; it gets weighted by some function $w(z, \tau)$, then renormalized, and what we get are observations from that weighted distribution. Now, very often, we are uncertain about the weight function.
The statistician's job is to model the weight function as well as the original or unrestricted underlying distribution. So I put parameters in both and see the symmetric role of theta and tau. Actually, I see also the symmetric role of the weight function and the underlying distribution. So, the Bayesian approach would assign a joint distribution to both theta and tau, the two parameters, and as we draw observations, we automatically learn about tau as we learn about theta. We learn about them in an equal even-handed way, even if we started with independent prior distributions on theta and tau. After a single observation, that independence would be gone because of the denominator and theta and tau become inextricably bound together. The non-Bayesian approach often takes the easy way out. It simply replaces tau by some estimate of tau and then proceeds to analyze the problem in terms of theta. But, as I say, the Bayesian approach provides a much more coherent view that brings into the foreground the symmetry between weight function and underlying density either as a weight function for the other and it no longer becomes fruitful even to think of them as weight functions and underlying densities.

Let me conclude with an aspect that we're currently working on rather hard and that is the question of information in weighted distributions that GP has touched on. Which experiment is more informative about this parameter theta that appears in the underlying density? A random sample unrestricted or unweighted if you like from the original distribution or a random sample from the weighted distribution.

The whole idea of encountered data, that we are forced to deal with these weighted distributions, comes out to be a sort of an unfortunate thing! If only we could get a random sample from the underlying distribution, life would be so simple!

Well, we're lucky at times that we have an encountered weighted distribution. We are often able to make much better inferences about theta from a sample from the weighted distribution than we are from a sample from the underlying distribution. Of course, we have to have the weight that is present in the encountered data. We have to have that modeled accurately.

Consider a very simple example of size biased sampling. If the underlying distribution is a gamma distribution, with parameters alpha and beta and if the weight function is of the size biased form, then the observation that we get with that size biased sampling would be simply a different gamma distribution with the alpha parameter increased by one. Thus, if the parameter alpha is known but the parameter beta is unknown in the underlying distribution, the size biased sample is more informative than the random sample from the original underlying distribution. So, instead of bemoaning the fact that we are stuck with a size biased sample, we should intentionally try to get a size biased sample. On the other hand, if it's the parameter alpha that is unknown and beta is known, then the situation reverses and the unrestricted random sample, unweighted random sample, is more informative than the sample from the weighted distribution. And if both parameters are known, this obviously indicates that neither experiment is more informative than the other in the strong sense and you have to look at what you are particularly interested in learning about in your own special problem.
REJOINER

G. P. Patil

It is a pleasure for me to be able to appreciate the thoughtful discussions presented by Drs. Joel O'Connor, N. Phillip Ross, and the late Morris H. DeGroot. The issues and approaches involving encountered data in statistical ecology and in environmental statistics are going to be of increasing importance in the days ahead. Their comments and suggestions should continue to be helpful to those interested.

I am also glad to report that, since the presentation of this paper, the American Statistical Association has initiated a new Section on Statistics and the Environment, the Ecological Society of America has initiated a new Section of Statistical Ecology, and a new International Environmetric Society has been launched sponsoring this timely journal called Environmetrics. The readers may also be interested to know that at the 1994 International Congress of Ecology, a Silver Jubilee Program on the Frontiers of Statistical Ecology and Ecological Statistics is under planning and discussion with an explicit theme of Statistical Ecology and Environmental Change. Let's be in touch.