Statistics in the Environmental Sciences

Gertz / London
editors

ASTM STP 845

STATISTICS IN THE ENVIRONMENTAL SCIENCES

A symposium
sponsored by ASTM
Committee D-19 on Water
Philadelphia, Pa., 7–8 October 1982

ASTM SPECIAL TECHNICAL PUBLICATION 845
Steven M. Gertz, Roy F. Weston, Inc.,
and M. D. London, Public Service Electric
and Gas Company, editors

ASTM Publication Code Number (PCN)
04-845000-16

1916 Race Street, Philadelphia, Pa. 19103

Library of Congress Cataloging in Publication Data

Statistics in the environmental sciences.

(ASTM special technical publication; 845)
Includes bibliographies and index.
1. Ecology—Statistical methods—Congresses.
2. Environmental monitoring—Statistical methods—
Congresses. 3. Environmental protection—Statistical
methods—Congresses. I. Gertz, Steven M. II. London,
M. D. III. American Society for Testing and Materials.
Committee D-19 on Water. IV. Series.
QH541.15.S72S73 1984 574.5′072 83-73439
ISBN 0-8031-0206-2

Library of Congress Catalog Card Number: 83-73439

NOTE

The Society is not responsible, as a body,
for the statements and opinions
advanced in this publication.

Printed in Baltimore, Md. (b)
November 1984
Second Printing, Baltimore, Md.
October 1986

Foreword

The symposium on Statistics in the Environmental Sciences was held in Philadelphia, Pennsylvania, 7-8 October 1982. The symposium was sponsored by ASTM Committee D-19 on Water. Steven M. Gertz, Roy F. Weston, Inc., presided as symposium chairman. Steven M. Gertz and M. D. London, Public Service Electric and Gas Company, are editors of this publication.

Related
ASTM Publications

Aquatic Toxicology and Hazard Assessment: Sixth Symposium, STP 802 (1983). 04-802000-16

Aquatic Toxicology and Hazard Assessment (Fifth Conference), STP 766 (1982), 04-766000-16

Native Aquatic Bacteria: Enumeration, Activity, and Ecology, STP 695 (1979), 04-695000-16

Methods and Measurements of Periphyton Communities: A Review, STP 690 (1978), 04-690000-16

Biological Data in Water Pollution Assessment: Quantitative and Statistical Analyses, STP 652 (1978), 04-652000-16

A Note of Appreciation to Reviewers

The quality of the papers that appear in this publication reflects not only the obvious efforts of the authors but also the unheralded, though essential, work of the reviewers. On behalf of ASTM we acknowledge with appreciation their dedication to high professional standards and their sacrifice of time and effort.

ASTM Committee on Publications

ASTM Editorial Staff

Contents

Introduction

Mathematical and statistical techniques are important tools for the scientist and engineer to use in describing a given environment, describing an existing impact, predicting an impact, or ascertaining an environmental trend. Historically, however, these techniques were not used frequently, and many environmental studies were more qualitative than quantitative. Yet, in recent years the need to develop and use more quantitative mathematical and statistical analyses has become apparent to many investigators. This need occurred for many reasons, such as: the desire to effectively and objectively communicate the results of an investigation; the desire to have a consistent measurable base upon which to establish environmental rules and regulations; the needs of industry to analyze and predict environmental impacts; the desire to monitor the environment such that that program cost is minimized and information content is maximized; and so on. In an effort to respond to these needs, ASTM Subcommittee D19.01 on Statistical Methods organized a symposium entitled "Statistics In the Environmental Sciences."

This symposium was held in Philadelphia, Pennsylvania on 7-8 October 1982, and was divided into two parts. Part 1 was at ASTM Headquarters on 7 October 1982 and consisted of the invited papers. The topics covered there and reported on in this symposium volume include:

1. Statistical Ecology
2. Environmental Monitoring
3. Ecological Impact Assessment
4. Risk Assessment
5. Environmental Health
6. Quality Assurance
7. Computer Software

Part 2 of the symposium was held on 8 October 1982 at the offices of Roy F. Weston, Inc., in West Chester, Pennsylvania. This segment began with a panel discussion on "A Proposed Cooperative Program on Statistical Ecology and Environmental Statistics" which is presented in this volume. The panel discussion was then followed by a series of workshops which examined in detail the topics presented in Part 1.

The purpose of this symposium and this publication is to provide a forum for scientists engaged in and scientists interested in learning more about the development and application of mathematical and statistical techniques to environmental problems. Emphasis was placed on the conceptual framework of applying statistical analyses but not at the expense of mathematical rigor. Rather, the approach taken emphasized the types of analyses available, their usefulness under a variety of conditions, and the decision making process an investigator must go through prior to applying the techniques.

The papers presented in this volume will hopefully be useful to those engaged in a variety of environmental studies. Overall, this volume should be particularly useful to the biologist, toxicologist, and environmental engineer since it provides guidance in the design of environmental studies programs and the evaluation and interpretation of subsequent data. It should also prove useful to regulatory personnel, industrial scientists and engineers, and professor and student alike.

Steven M. Gertz

Roy F. Weston, Inc., West Chester, Pa. 19380.

G. P. Patil[1]

Some Perspectives of Statistical Ecology and Environmental Statistics*

REFERENCE: Patil, G. P., **"Some Perspectives of Statistical Ecology and Environmental Statistics,"** *Statistics in the Environmental Sciences, ASTM STP 845*, S. M. Gertz and M. D. London, Eds., American Society for Testing and Materials, 1984, pp. 3–22.

ABSTRACT: The paper sets the scene with historical and professional organizational perspectives briefly describing the activities and publications of groups concerned with statistical ecology and environmental statistics. Then follow the scientific perspectives with examples illustrating certain basic issues and features arising in ecological and environmental work. The observational studies, visibility bias, and inferential recovery constitute a critical problem area. It is discussed at some length with examples and data sets. Extrapolation issues constitute another critical problem area. Different kinds of extrapolation are briefly discussed with examples of low-dose extrapolation and calculated virtually safe dose levels. An example of recruitment distributions and inferences about long-term yields is also given. The issue of single versus several models is touched upon. The present day status of the subject of quantitative risk analysis is briefly described. A need for a focal cooperative program in statistical ecology and environmental statistics is identified.

KEY WORDS: statistical ecology, environmental statistics, cooperative program, size-biased sampling, weighted distributions, observational bias, extrapolation issues, risk analysis, recruitment distributions, virtual safe dose levels

The third quarter of the twentieth century proceeded with a sense of achievement in matters of productivity and prosperity and dreamed of leisure and enriched life in the near future. The fourth quarter of the twentieth century has seen only nightmares so far! Within this complex mosaic of science, technol-

*Based on the keynote speech given at the inaugural of the ASTM Symposium on Statistics and the Environmental Sciences. This paper was prepared while the author was under partial support from the National Marine Fisheries Service on a project entitled, Stochastics and Statistics in Marine Fisheries Research and Management, under contract NA-80-FAC-00040.

[1]Professor of Mathematical Statistics, Department of Statistics and the Graduate Program in Ecology, The Pennsylvania State University, University Park, Pa. 16802.

ogy, and the society, we need to be pursuing statistics, ecology, and the environment with a firm and full view of productivity, prosperity, and quality. I would like to share with you some of my perspectives of statistical ecology and environmental statistics—a subject area that has acquired importance in view of the current ecological and environmental issues.

The subject area of statistical ecology and environmental statistics is relatively young. Various interdisciplinary efforts and activities by concerned professional organizations have been instrumental in the promotion of scientific dialogues and in the dissemination of the results in an impressive variety of valuable publications. Among others, the following organizations and groups come to mind in this connection: the Committee on Statistics and the Environment of the American Statistical Association, the SIAM Institute for Mathematics and Society, the Mathematical Ecology Group in Great Britain, and the International Statistical Ecology Program of the International Association for Ecology, the International Statistical Institute, and the International Biometric Society. Broadly speaking, the latter two groups have so far emphasized statistical ecology in their programs, whereas the former have had primary emphasis on environmental statistics.

Publications listed at the end of this paper can give an idea of these timely and useful volumes [*1–24*]. They bring home the underlying unity in diversity while effectively making available in one place the enormous material of an interdisciplinary nature. This is what most everyone would like to know. Having these volumes around is extremely worthwhile.

Most everyone would like to know of the proposed cooperative program on statistical ecology and environmental statistics for increased resource productivity and environmental protection. The present symposium has a full scale panel discussion on the proposed program, and a special paper prepared for the purposes of this volume. It would give an idea of what could be done together to meet the challenge of statistics, ecology, and the environment.

Certain Scientific Perspectives Through Examples

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

Comprehensive Versus Comprehensible

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

1. For lack of information, we do not (quite) comprehend the situation.

2. We collect information, and we tend to collect comprehensive information.

3. Because the information is comprehensive, we do not (quite) comprehend it.

4. So we summarize the information through a set of indices (statistics) so that it would be comprehensible.
5. But now, we do not comprehend quite what the indices exactly mean.
6. And, therefore, we do not (quite) comprehend the situation.
7. 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 for 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.

Information Paradox

Generally speaking, we have infinite information around us; but often we do not quite know what to make of it. At the same time, however, for a specific problem, we may have an infinitesimal amount of information. Often, we do not quite know what to make of it either. We find ourselves in need of additional information, even when we know of the infinite information around us.

Is there a satisfactory resolution of the paradox?

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 375 individual members of the species under consideration. The judge invites comments.

Industrial Lobby—The reported record of 375 members makes sense. The visibility factor is low in such surveys. The investigator has surely missed most of them that are out there. The exploitation should not cause alarm.

Environmental Lobby—The reported record of 375 makes sense. The investigator is an expert in such surveys. He has observed and recorded most of them that are there. And, therefore, only a few are out there. The species population needs to be protected.

The scenario is a typical one. It brings home the issues characteristic of field observations often lacking a sampling frame necessary for the classical sampling theory to apply. One needs to work with visibility analysis instead. Satisfactory estimation of biological population abundance depends, in such cases, largely on adequate measurement of visibility, variously termed catchability, audibility, etc. And, this is not a trivial problem!

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, stands facing north, and observes a bird vanish at the horizon at an angle of 10 deg. A little later he finds a second bird vanish at the horizon at an angle of 350 deg. He obtains the average $(10+350)/2 = 180$ deg and declares that south is the typical direction. Figure 1, however, reveals that this 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 do not mismatch or contradict the basics of the protocol of the phenomenon that is being modeled.

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.

Observational Studies, Visibility Bias, and Inferential Recovery with Examples

Ecological surveys and environmental studies, unlike most of the socioeconomic survey work, frequently lack a sampling frame for the populations under study. Observational mechanisms become procedures of unequal probability sampling without a sampling frame for reasons of differential visibility, audibility, or catchability. The original distribution produced by nature, thus, may not be reproduced when an investigator collects a sample of observations.

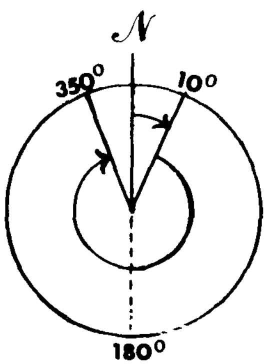

FIG. 1—*North is a typical direction.*

The main interest in any investigation is, however, to determine the characteristics of the original distribution. Further, it also becomes important to assess the distortion caused in determining these characteristics in case the change in the underlying distribution due to sampling bias is ignored. Some of these issues and examples are briefly discussed. A demonstration project on sex ratio is given along with the analysis of the data collected at the Symposium. The observational studies aspect in statistical ecology and environmental statistics is a critical problem area needing careful attention and effort.

Size-Biased Sampling and Weighted Distributions

Size-biased sampling arises in ecological and environmental work quite often. A common feature in the discrete case is that individuals are sampled, whereas aggregate observations are recorded on groups to which the individuals belong, such as, herds, families, species, etc. In the continuous case, a common feature is that the sampling unit for which a measurement is recorded is not selected at random from the population of units, but through a built-in mechanism that makes the selection probability proportional to the recorded measurement of the unit.

Suppose X is a nonnegative observable random variable (rv) with its natural probability density function $f(x;\ \theta)$, where θ is the natural parameter. Suppose a realization x of X under $f(x;\ \theta)$ enters the investigator's record with probability proportional to $w(x,\ \beta)$, so that

$$\begin{aligned}\text{prob}(\text{recording}|X = y) &\div \text{prob}(\text{recording}|X = x)\\ &= w(y,\ \beta) \div w(x,\ \beta)\end{aligned}$$

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

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

where ω is the normalizing factor. The rv X^w is called the weighted version of X, and its distribution is called the weighted distribution with weight function w. When $w(x,\ \beta) = x$, the weighted distribution is called the size-biased distribution with *pdf*

$$f^*(x;\theta) = \frac{xf(x;\theta)}{\mu}, \quad \mu = E[X]$$

and the corresponding sighting mechanism is called size-biased sampling. For further discussion and references, see Patil and Rao [*25,26*].

Aerial Survey and Visibility Bias

The visibility bias is a recognized problem in aerial survey techniques for estimating wildlife population density and for studying animal behavior. The visibility bias is present because of the failure to observe some animals. A sampling model requires: (*a*) animals occur in groups of varying sizes, (*b*) each animal has probability β of being observed, and (*c*) conditional on observing at least one member of a group, the entire group is observed with certainty. A parameter of interest may be the mean group size. If the group size X has *pdf* $f(x;\ \theta)$, then, under the above assumptions, the distribution of the recorded group size X^w has *pdf* $f^w(x;\ \theta,\ \beta)$ with weight function $w(x,\ \beta) = 1 - (1 - \beta)^x$. As $\beta \to 0$, the limit of $f^w(x;\ \theta,\ \beta)$ assumes the size-biased form $f^*(x;\ \theta)$.

For further discussion of the problems and available solutions, see Cook and Martin [*27*] and Patil and Rao [*25,26*].

Resource Utilization Surveys

In assessing the extent of utilization of the national parks, tourist hotels, and other recreational facilities, an investigator asks a tourist how long he has been visiting the location. The data so recorded on a sample of tourists have the size-biased feature. This is because tourists with longer duration visits come into the record with larger probabilities than those with shorter duration visits. It is considered reasonable to assume the encounter probability to be proportional to the duration.

One could avoid introducing such a size-bias by asking the question to each departing individual instead, at the exit. But, then, this would require the investigator to invest considerable survey time.

For further discussion of the problems and available solutions, see Mahfoud, Patil, and Ratnaparkhi [*28*], Morrison [*29*], and Patil and Rao [*25*].

In 1966, a survey research agency undertook a statistical survey on tourism in a certain country. One of the main objectives of this investigation was to estimate the mean sojourn time per tourist. Due to the lack of a sampling frame and the high mobility of the tourists, it was decided to sample them where they could be easily reached. So two types of surveys were conducted simultaneously. In one survey, the tourists were sampled in the hotels where they were residing, and in the other survey, the sample was selected at different frontier stations from tourists who were about to end their journey in the country. In both cases, it was assumed that the selection procedure adopted would produce random samples. It turned out, however, that the observed mean sojourn time ($\mu = 17.8$ days per tourist) obtained from the survey in the hotels was practically twice the observed mean sojourn time ($\mu = 9.0$ days per tourist) computed from the sample obtained at the frontier stations.

No explanation was given as to what might have caused these two estimates to differ so much. Both estimates were based on large sample sizes: 12 321 tourists from the frontier stations and 3 000 from the hotels. Suspected by the officials in the Department of Planning, the estimate from the hotels was discarded.

It was overlooked then that the sampling scheme in the hotels did not lead to a random sample of observations on the sojourn time because it gave larger probability of selection to those tourists with longer residence time in the hotels. If this phenomenon were realized, an adequate estimator of the mean sojourn time would have been developed, thus preventing the sacrifice of 20% of the total collected information.

Trait of Interest and Family Size

If we wish to study the distribution of albino children in families capable of producing such children, we may contact a large number of families and ascertain from each family the number of albinos. But, this method of investigation is wasteful. A convenient method in such a case is first to discover an albino child and through it obtain the information about the family to which the child belongs. But such a procedure may not give equal chance to all families in which albinos have occurred. The exact chance for a family with x albinos is that of detecting at least one of its albino children, which may be equal to $w(x, \beta) = 1 - (1 - \beta)^x$, where β is the probability of an encounter with an albino child. As $\beta \to 0$, the limit of $f^w(x: \theta, \beta)$ assumes the size-biased form $f^*(x; \theta)$.

For further discussion of the problems and available solutions, see Patil and Rao [*26*] and Stene [*30*].

Family Size and Sex Ratio

Various demographic and social studies involve family size and sex ratio as important factors which have some bearing on main study. Let us ascertain from each male student in a class the number of brothers, including himself, and the sisters he has. Denote by k the number of reporting male students, and by B and S the total number of brothers and sisters reported. The characteristics of the population of interest are: (a) the male sex ratio, and (b) the mean family size (number of children).

The data in Table 1 are from Patil and Rao [*26*]. For each city indicated, it gives the number k of male respondents with the number of brothers B and the number of sisters S reported.

Under the usual binomial model, $B/(B + S)$ would give the unbiased estimate of the male sex ratio, whereas, $(B + S)/k$ would estimate the mean family size. However, under size-biased sampling, these become over-estimates. Corrected estimates for the sex ratio are $(B - k)/(B + S - k)$, whereas the corrected estimates for the mean family size are given by the harmonic means (δ) of the reported family sizes for each city.

We note that the values of $(B + k)/(B + S - k)$ are closer to 1/2, and that