United States
Environmental Protection
Agency

Office of Research and
Development
Washington DC 20460

EPA/630/R-98/004
January 1999

EPA

Report of the Workshop on Selecting Input Distributions for Probabilistic Assessments

RISK ASSESSMENT FORUM

EPA/630/R-98/004
January 1999

Report of the Workshop on Selecting Input Distributions For Probabilistic Assessments

U.S. Environmental Protection Agency
New York, NY
April 21-22, 1998

Risk Assessment Forum
U.S. Environmental Protection Agency
Washington, DC 20460

NOTICE

This document has been reviewed in accordance with U.S. Environmental Protection Agency (EPA) policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

This report was prepared by Eastern Research Group, Inc. (ERG), an EPA contractor (Contract No. 68-D5-0028, Work Assignment No. 98-06) as a general record of discussions during the Workshop on Selecting Input Distributions for Probabilistic Assessments. As requested by EPA, this report captures the main points and highlights of discussions held during plenary sessions. The report is not a complete record of all details discussed nor does it embellish, interpret, or enlarge upon matters that were incomplete or unclear. Statements represent the individual views of each workshop participant; none of the statements represent analyses by or positions of the Risk Assessment Forum or the EPA.

CONTENTS

CONTENTS (Continued)

APPENDICES

SECTION ONE

INTRODUCTION

1.1 BACKGROUND AND PURPOSE

The U.S. Environmental Protection Agency (EPA) has long emphasized the importance of adequately characterizing uncertainty and variability in its risk assessments, and it continuously studies various quantitative techniques for better characterizing uncertainty and variability. Historically, Agency risk assessments have been deterministic (i.e., based on a point estimate), and uncertainty analyses have been largely qualitative. In May 1997, the Agency issued a policy on the use of probabilistic techniques in characterizing uncertainty and variability. This policy recognizes that probabilistic analysis tools like Monte Carlo analysis are acceptable provided that risk assessors present adequate supporting data and credible assumptions. The policy also identifies several implementation activities that are designed to help Agency assessors review and prepare probabilistic assessments.

To this end, EPA's Risk Assessment Forum (RAF) is developing a framework for selecting input distributions for probabilistic assessment. This framework emphasizes parametric distributions, estimations of the parameters of candidate distributions, and evaluations of the candidate distributions' quality of fit. A technical panel, convened under the auspices of the RAF, began work on the framework in the summer of 1997. In September 1997, EPA sought input on the framework from 12 experts from outside the Agency. The group's recommendations included:

- Expanding the framework's discussion of exploratory data analysis and graphical methods for assess the quality of fit.
- Discussing distinctions between variability and uncertainty and their implications.
- Discussing empirical distributions and bootstrapping.
- Discussing correlation and its implications.
- Making the framework available to the risk assessment community as soon as possible.

In response to this input, EPA initiated a pilot program in which the Research Triangle Institute (RTI) applied the framework for fitting distributions to data from EPA's Exposure Factors Handbook (EFH) (US EPA, 1996a). RTI used three exposure factors—drinking water intake, inhalation rate, and residence time—as test cases. Issues highlighted as part of this effort fall into two broad categories: (1) issues associated with the *representativeness* of the data, and (2) issues associated with using the *Empirical Distribution Function (EDF)* (or resampling techniques) versus using a theoretical *Parametric Distribution Function (PDF)*.

In April 1998, the RAF organized a 2-day workshop, "Selecting Input Distributions for Probabilistic Assessments," to solicit expert input on these and related issues. Specific workshop goals included:

- Discussing issues associated with the selection of probability distributions.

- Obtaining expert input on measurements, extrapolations, and adjustments.

- Discussing qualitatively how to make quantitative adjustments.

EPA developed two issue papers to serve as a focal point for discussions: "Evaluating Representativeness of Exposure Factors Data" and "Empirical Distribution Functions and Non-parametric Simulation." These papers which were developed strictly to prompt discussions during the workshop are found in Appendix A. Discussions during the 2-day workshop focused on technical issues, not policy. The experts discussed issues that would apply to any exposure data.

This workshop report is intended to serve as an information piece for Agency assessors who prepare or review assessments based on the use of probabilistic techniques and who work with various exposure data. This report does not represent Agency guidance. It simply attempts to capture the technical rigor of the workshop discussions and will be used to support further development and application of probabilistic analysis techniques/approaches.

1.2 WORKSHOP ORGANIZATION

The workshop was held on April 21 and 22, 1998, at the EPA Region 2 offices in New York City. The 21 participants, experts in exposure and risk assessment, included biologists, chemists, engineers, mathematicians, physicists, statisticians, and toxicologists, and represented industry, academia, state agencies, EPA, and other federal agencies. A limited number of observers also attended the workshop. The experts and observers are listed in Appendix B.

The workshop agenda is in Appendix C. Mr. McCabe (EPA Region 2), Steven Knott of the RAF, and Dr. H. Christopher Frey, workshop facilitator, provided opening remarks. Before discussions began, Ms. Jacqueline Moya and Dr. Timothy Barry of EPA summarized the two issue papers.

During the 2-day workshop, the technical experts exchanged ideas in plenary and four small group breakout sessions. Discussions centered on the two issue papers distributed for review and comment before the workshop. Detailed discussions focused primarily on the questions in the charge (Appendix D). "Brainwriting" sessions were held within the smaller groups. Brainwriting, an interactive technique, enabled the experts to document their thoughts on a topic and build on each others' ideas. Each small group captured the essence of these sessions and presented the main ideas to the entire group during plenary sessions. A compilation of notes from the breakout sessions are included in Appendix E. Following expert input, observers were allowed to address the panel with questions or comments. In addition to providing input at the workshop, several experts provided pre- and postmeeting comments, which are in Appendices F and G, respectively.

Section Two of this report contains the chairperson's summary of the workshop. Section Three highlights workshop opening remarks. Section Four summarizes Agency presentations of the two issue papers. Sections Five and Six describe expert input on the two main topic areas—representativeness and EDF/PDF issues. Speakers' presentation materials (overheads and supporting papers) are included in Appendix H.

SECTION TWO

CHAIRPERSON'S SUMMARY
Prepared by: H. Christopher Frey, Ph.D.

The workshop was comprised of five major sessions, three of which were devoted to the issue of representativeness and two to issues regarding parametric versus empirical distributions and goodness-of-fit. Each session began with a trigger question. For the three sessions on representativeness, there was discussion in a plenary setting, as well as discussions within four breakout groups. For the two sessions regarding selection of parametric versus empirical distributions and the use of goodness-of-fit tests, the discussions were conducted in plenary sessions.

2.1 REPRESENTATIVENESS

The first session covered three main questions, based on the portion of the workshop charge (Appendix D) requesting feedback on the representativeness issue paper. After some general discussion, the following three trigger questions were formulated and posed to the group:

1. What information is required to fully specify a problem definition?
2. What constitutes (lack of) representativeness?
3. What considerations should be included in, added to, or excluded from the checklists given in the issue paper on representativeness (Appendix A)?

The group was then divided into four breakout groups, each of which addressed all three of these questions. Each group was asked to use an approach known as "brainwriting." Brainwriting is intended to be a silent activity in which each member of a group at any given time puts thoughts down on paper in response to a trigger question. After completing an idea, a group member exchanges papers with another group member. Typically, upon reading what others have written, new ideas are generated and written down. Thus, each person has a chance to read and respond to what others have written. The advantages of brainwriting are that all participants can generate ideas simultaneously, there is less of a problem with domination of the discussion by just a few people, and a written record is produced as part of the process. A disadvantage is that there is less "interaction" with the entire group. After the brainwriting activity was completed, a representative of each group reported the main ideas to the entire group.

The experts generally agreed that before addressing the issue of representativeness, it is necessary to have a clear problem definition. Therefore, there was considerable discussion of what factors must be considered to ensure a complete problem definition. The most general requirement for a good problem definition, to which the group gave general assent, is to specify the "who, what, when, where, why, and how." The "who" addresses the population of interest. "Where" addresses the spatial characteristics of the assessment. "When" addresses the temporal characteristics of the assessment. "What" relates to the specific chemicals and health effects of concern. "Why" and "how" may help clarify the previous matters. For example, it is helpful to know that exposures occur because of a particular behavior (e.g., fish consumption) when attempting to define an exposed population and the spatial and temporal extent of the problem. Knowledge of "why" and "how" is also useful later for proposing mitigation or prevention strategies. The group in general agreed upon these principles for a problem definition, as well as the more specific suggestions detailed in Section 5.1.1 of this workshop report.

In regard to the second trigger question, the group generally agreed that "representativeness" is context-specific. Furthermore, there was a general trend toward finding other terminology instead of using the term "representativeness." In particular, many the group concurred that an objective in an assessment is to make sure that it is "useful and informative" or "adequate" for the purpose at hand. The adequacy of an assessment may be evaluated with respect to considerations such as "allowable error" as well as practical matters such as the ability to make measurements that are reasonably free of major errors or to reasonably interpret information from other sources that are used as an input to an assessment. Adequacy may be quantified, in principle, in terms of the precision and accuracy of model inputs and model outputs. There was some discussion of how the distinction between variability and uncertainty relates to assessment of adequacy. For example, one may wish to have accurate predictions of exposures for more than one percentile of the population, reflecting variability. For any given percentile of the population, however, there may be uncertainty in the predictions of exposures. Some individuals pointed out that, because often it is not possible to fully validate many exposure predictions or to obtain input information that is free of error or uncertainty, there is an inherently subjective element in assessing adequacy. The stringency of the requirement for adequacy will depend on the purpose of the assessment. It was noted, for example, that it may typically be easier to adequately define mean values of exposure than upper percentile values of exposure. Adequacy is also a function of the level of detail of an assessment; the requirements for adequacy of an initial, screening-level calculation will typically be less rigorous than those for a more detailed analysis.

Regarding the third trigger question, the group was generally complimentary of the proposed checklists in the representativeness issue paper (see Appendix A). The group, however, had many suggestions for improving the checklists. Some of the broader concerns were about how to make the checklists context-specific, because the degree of usefulness of information depends on both the quality of the information and the purpose of the assessment. Some of the specific suggestions included using flowcharts rather than lists; avoiding overlap among the flowcharts or lists; developing an interactive Web-based flowchart that would be flexible and context-specific; and clarifying terms used in the issue paper (e.g., "external" versus "internal" distinction). The experts also suggested that the checklists or flowcharts encourage additional data collection where appropriate and promote a "value of information" approach to help prioritize additional data collection. Further discussion of the group's comments is given in Section 5.1.3.

2.2 SENSITIVITY ANALYSIS

The second session was devoted to issues encapsulated in the following trigger questions:

> How can one do sensitivity analysis to evaluate the implications of non-representativeness? In other words, how do we assess the importance of non-representativeness?

The experts were asked to consider data, models, and methods in answering these questions. Furthermore, the group was asked to keep in mind that the charge requested recommendations for immediate, short-term, and long-term studies or activities that could be done to provide methods or examples for answering these questions.

There were a variety of answers to these questions. A number of individuals shared the view that non-representativeness may not be important in many assessments. Specifically, they argued that many assessments and decisions consider a range of scenarios and populations. Furthermore, populations and exposure scenarios typically change over time, so that if one were to focus on making an assessment "representative" for one point in time or space, it could fail to be representative at other points in time or space or even for the original population of interest as individuals enter, leave, or change within the exposed population. Here again the notion of adequacy, rather than representativeness, was of concern to the group.

The group reiterated that representativeness is context-specific. Furthermore, there was some discussion of situations in which data are collected for "blue chip" distributions that are not specific to any particular decision. The experts did recommend that, in situations where there may be a lack of adequacy of model predictions based on available information, the sensitivity of decisions should be evaluated under a range of plausible adjustments to the input assumptions. It was suggested that there may be multiple tiers of analyses, each with a corresponding degree of effort and rigor regarding sensitivity analyses. In a "first-tier" analysis, the use of bounding estimates may be sufficient to establish sensitivity of model predictions with respect to one or more model outputs, without need for a probabilistic analysis. After a preliminary identification of sensitive model inputs, the next step would typically be to develop a probability distribution to represent a plausible range of outcomes for each of the sensitive inputs. Key questions to be considered are whether to attempt to make adjustments to improve the adequacy or representativeness of the assumptions and/or whether to collect additional data to improve the characterization of the input assumptions.

One potentially helpful criterion for deciding whether data are adequate is to try to answer the question: "Are the data good enough to replace an assumption?" If not, then additional data collection is likely to be needed. One would need to assess whether the needed data can be collected. A "value of information" approach can be useful in prioritizing data collection and in determining when sufficient data have been collected.

There was some discussion of sensitivity analysis of uncertainty versus sensitivity analysis of variability. The experts generally agreed that sensitivity analysis to identify key sources of uncertainty is a useful and appropriate thing to do. There was disagreement among the experts regarding the meaning of identifying key sources of variability. One expert argued that identifying key sources of variability is not useful, because variability is irreducible. However, knowledge of key sources of variability can be useful in identifying key characteristics of highly exposed subpopulations or in formulating prevention or mitigation measures. Currently, there are many methods that exist for doing sensitivity analysis, including running models for alternative scenarios and input assumptions and the use of regression or statistical methods to identify the most sensitive input distributions in a probabilistic analysis. In the short-term and long-term, it was suggested that some efforts be devoted to the development of "blue chip" distributions for quantities that are widely used in many exposure assessments (e.g., intake rates of various foods). It was also suggested that new methods for sensitivity analysis might be obtained from other fields, with specific examples based on classification schemes, time series, and "g-estimation."

2.3 MAKING ADJUSTMENTS TO IMPROVE REPRESENTATION

In the third session, the group responded to the following trigger question:

How can one make adjustments from the sample to better represent the population of interest?

The group was asked to consider "population," spatial, and temporal characteristics when considering issues of representativeness and methods for making adjustments. The group was asked to provide input regarding exemplary methods and information sources that are available now to help in making such adjustments, as well as to consider short-term and long-term research needs.

The group clarified some of the terminology that was used in the issue paper and in the discussions. The term "population" was defined as referring to "an identifiable group of people." The experts noted that often one has a sample of data from a "surrogate population," which is not identical to the "target population" of interest in a particular exposure assessment. The experts also noted that there is a difference between the "analysis" of actual data pertaining to the target population and "extrapolation" of information from data for a surrogate population to make inferences regarding a target population. It was noted that extrapolation always "introduces" uncertainty.

On the temporal dimension, the experts noted that, when data are collected at one point in time and are used in an assessment aimed at a different point in time, a potential problem may occur because of shifts in the characteristics of populations between the two periods.

Reweighting of data was one approach that was mentioned in the plenary discussion. There was a discussion of "general" versus mechanistic approaches for making adjustments. The distinction here was that "general" approaches might be statistical, mathematical, or empirical in their foundations (e.g., regression analysis), whereas mechanistic approaches would rely on theory specific to a particular problem area (e.g., a physical, biological, or chemical model). It was noted that temporal and spatial issues are often problem-specific, which makes it difficult to recommend universal approaches for making adjustments. The group generally agreed that it is desirable to include or state the uncertainties associated with extrapolations. Several participants strongly expressed the view that "it is okay to state what you don't know," and there was no disagreement on this point.

The group recommended that the basis for making any adjustments to assumptions regarding populations should be predicated on stakeholder input and the examination of covariates. The group noted that methods for analyzing spatial and temporal aspects exist, if data exist. Of course, a common problem is scarcity of data and a subsequent reliance on surrogate information. For assessment of spatial variations, methods such as kreiging and random fields were commonly suggested. For assessment of temporal variations, time series methods were suggested.

There was a lively discussion regarding whether adjustments should be "conservative." Some experts initially argued that, to protect public health, any adjustments to input assumptions should tend to be biased in a conservative manner (so as not to make an error of understating a health risk, but with some nonzero probability of making an error of overstating a particular risk). After some additional discussion, it appeared that the experts were in agreement that one should strive primarily for accuracy and that ideally any adjustments that introduce "conservatism" should be left to decision makers. It was pointed out that invariably many judgments go into the development of input assumptions for an analysis and that these judgments in reality often introduce some conservatism. Several pointed out that "conservatism" can entail significant costs if it results in over control or misidentification of important risks. Thus, conservatism in individual assessments may not be optimal or even conservative in a broader sense if some sources of risk are not addressed because others receive undue attention. Therefore, the overall recommendation of the experts regarding this issue is to strive for accuracy rather than conservatism, leaving the latter as an explicit policy issue for decision makers to introduce, although it is clear that individual participants had somewhat differing views.

The group's recommendations regarding measures that can be taken now include the use of stratification to try to reduce variability and correlation among inputs in an assessment, brainstorming to generate ideas regarding possible adjustments that might be made to input assumptions, and stakeholder input for much the same purpose, as well as to make sure that no significant pathways or scenarios have been overlooked. It was agreed that "plausible extrapolations" are reasonable when making adjustments to improve representativeness or adequacy. What is "plausible" will be context-specific.

In the short term, the experts recommended that the following activities be conducted:

Numerical Experiments. Numerical experiments can be used to test existing and new methods for making adjustments based on factors such as averaging times or averaging areas. For example, the precision and accuracy of the Duan-Wallace model (described in the representativeness issue paper in Appendix A) for making adjustments from one averaging time to another can be evaluated under a variety of conditions via numerical experiments.

Workshop on Adjustment Methods. The experts agreed in general that there are many potentially useful methods for analysis and adjustment but that many of these are to be found in fields outside the risk analysis community. Therefore, it would be useful to convene a panel of experts from other fields for the purpose of cross-disciplinary exchange of information regarding methods applicable to risk analysis problems. For example, it was suggested that geostatistical methods should be investigated.

Put Data on the Web. There was a fervent plea from at least one expert that data for "blue chip" and other commonly used distributions be placed on the Web to facilitate the dissemination and analysis of such data. A common concern is that often data are reported in summary form, which makes it difficult to analyze the data (e.g., to fit distributions). Thus, the recommendation includes the placement of actual data points, and not just summary data, on publicly accessible Web sites.

Suggestions on How to Choose a Method. The group felt that, because of the potentially large number of methods and the need for input from people in other fields, it was unrealistic to provide recommendations regarding specific methods for making adjustments. However, they did suggest that it would be possible to create a set of criteria regarding desirable features for such methods that could help an assessor when making choices among many options.

In the longer term, the experts recommend that efforts be directed at more data collection, such as improved national or regional surveys, to better capture variability as a function of different populations, locations, and averaging times. Along these lines, specific studies could be focused on the development or refinement of a select set of "blue chip" distributions, as well as targeted at updating or extending existing data sets to improve their flexibility for use in assessments of various populations, locations, and averaging times. The group also noted that because populations, pathways, and scenarios change over time, there will be a continuing need to improve existing data sets.

2.4 EMPIRICAL AND PARAMETRIC DISTRIBUTION FUNCTIONS

In the fourth session, the experts began to address the second main set of issues as given in the charge. The trigger question used to start the discussion was:

> What are the primary considerations in choosing between the use of parametric distribution functions (PDFs) and empirical distribution functions (EDFs)?

The group was asked to consider the advantages of using one versus the other, whether the choice is merely a matter of preference, whether one is preferred, and whether there are cases when neither should be used.

The initial discussion involved clarification of the difference between the terms EDF and "bootstrap." Bootstrap simulation is a general technique for estimating confidence intervals and characterizing sampling distributions for statistics, as described by Efron and Tibshirani (1993). An EDF can be described as a stepwise cumulative distribution function or as a probability density function in which each data point is assigned an equal probability. Non-parametric bootstrap can be used to quantify sampling distributions or confidence intervals for statistics based upon the EDF, such as percentiles or moments. Parametric bootstrap methods can be used to quantify sampling distributions or confidence intervals for statistics based on PDFs. Bootstrap methods are also often referred to as "resampling" methods. However, "bootstrap" and EDF are not the same thing.

The experts generally agreed that the choice of EDF versus PDF is usually a matter of preference, and they also expressed the general opinion that there should be no rigid guidance requiring the use of one or the other in any particular situation. The group briefly addressed the notion of consistency. While consistency in the use of a particular method (e.g., EDF or PDF in this case) may offer benefits in terms of simplifying analyses and helping decision makers, there was a concern that any strict enforcement of consistency will inhibit the development of new methods or the acquisition of new data and may also lead to compromises from better approaches that are context-specific. Here again, it is important to point out that the experts explicitly chose not to recommend the use of either EDF or PDF as a single preferred approach but rather to recommend that this choice be left to the discretion of assessors on a case-by-case basis. For example, it could be reasonable for an assessor to include EDFs for some inputs and PDFs for others even within the same analysis.

Some participants gave examples of situations in which they might prefer to use an EDF, such as: (a) when there are a large number of data points (e.g., 12,000); (b) access to high speed data storage and retrieval systems; (c) when there is no theoretical basis for selecting a PDF; and/or (d) when one has an "ideal" sample. There was some discussion of preference for use of EDFs in "data-rich" situations rather than "data-poor" situations. However, it was noted that "data poor" is context-specific. For example, a data set may be adequate for estimating the 90th percentile but not the 99th percentile. Therefore, one may be "data rich" in the former case and "data poor" in the latter case with the same data set.

Some experts also gave examples of when they would prefer to use PDFs. A potential limitation of conventional EDFs is that they are restricted to the range of observed data. In contrast, PDFs typically intuitive or theoretical appeal. PDFs are also preferred by some because they provide a compact representation of data and can provide insight into generalizable features of a data set. Thus, in contrast to the proponent of the use of an EDF for a data set of 12,000, another expert suggested it would be easier to summarize the data with a PDF, as long as the fit was reasonable. At least one person suggested that a PDF may be easier to defend in a legal setting, although there was no consensus on this point.

For both EDFs and PDFs, the issue of extrapolation beyond the range of observed data received considerable discussion. One expert stated that, the "further we go out in the tails, the less we know," to which another responded, "when we go beyond the data, we know nothing." As a rebuttal, a third expert asked "do we really know nothing beyond the maximum data point?" and suggested that analogies with similar situations may provide a basis for judgments regarding extrapolation beyond the observed data. Overall, most or all of the experts appeared to support some approach to extrapolation beyond observed data, regardless of whether one prefers an EDF or a PDF. Some argued that one has more control over extrapolations with EDFs, because there are a variety of functional forms that can be appended to create a "tail" beyond the range of observed data. Examples of these are described in the issue paper. Others argued that when there is a theoretical basis for selecting a PDF, there is also some theoretical basis for extrapolating beyond the observed data. It was pointed out that one should not always focus on the "upper" tail; sometimes the lower tail of a model input may lead to extreme values of a model output (e.g., such as when an input appears in a denominator).

There was some discussion of situations in which neither an EDF or a PDF may be particularly desirable. One suggestion was that there may be situations in which explicit enumeration of all combinations of observed data values for all model inputs, as opposed to a probabilistic resampling scheme, may be desired. Such an approach can help, for example, in tracing combinations of input values that produce extreme values in model outputs. One expert suggested that neither EDFs nor PDFs are useful when there must be large extrapolations into the tails of the distributions.

A question that the group chose to address was, "How much information do we lose in the tails of a model output by not knowing the tails of the model inputs?" One comment was that it may not be necessary to accurately characterize the tails of all model inputs because the tails (or extreme values) of model outputs may depend on a variety of other combinations of model input values. Thus, it is possible that even if no effort is made to extrapolate beyond the range of observed data in model inputs, one may still predict extreme values in the model outputs. The use of scenario analysis was suggested as an alternative or supplement to probabilistic analysis in situations in which either a particular input cannot reasonably be assigned a probability distribution or when it may be difficult to estimate the tails of an important input distribution. In the latter case, alternative upper bounds on the distribution, or alternative assumptions regarding extrapolation to the tails, should be considered as scenarios.

Uncertainty in EDFs and PDFs was discussed. Techniques for estimating uncertainties in the statistics (e.g., percentiles) of various distributions, such as bootstrap simulation, are available. An example was presented for a data set of nine measurements, illustrating how the uncertainty in the fit of a parametric distribution was greatest at the tails. It was pointed out that when considering alternative PDFs (e.g., Lognormal vs. Gamma) the range of uncertainty in the upper percentiles of the alternative distributions will typically overlap; therefore, apparent differences in the fit of the tails may not be particularly significant from a statistical perspective. Such insights are obtained from an explicit approach to distinguishing between variability and uncertainty in a "two-dimensional" probabilistic framework.

The group discussed whether mixture distributions are useful. Some experts were clearly proponents of using mixture distributions. A few individuals offered some cautions that it can be difficult to know when to properly employ mixtures. One example mentioned was for radon concentrations. One expert mentioned in passing that radon concentrations had been addressed in a particular assessment assuming a Lognormal distribution. Another responded that the concentration may more appropriately be described as a mixture of normal distributions. There was no firm consensus on whether it is better to use a mixture of distributions as opposed to a "generalized" distribution that can take on many arbitrary shapes. Those who expressed opinions tended to prefer the use of mixtures because they could offer more insight about processes that produced the data.

Truncation of the tails of a PDF was discussed. Most of the experts seemed to view this as a last resort fraught with imperfections. The need for truncation may be the result of an inappropriate selection of a PDF. For example, one participant asked, "If you truncate a Lognormal, does this invalidate your justification of the Lognormal?" It was suggested that alternative PDFs (perhaps ones that are less "tail heavy") be explored. Some suggested that truncation is often unnecessary. Depending upon the probability mass of the portion of the distribution that is considered for truncation, the probability of sampling an extreme value beyond a plausible upper bound may be so low that it does not occur in a typical Monte Carlo simulation of only a few thousand iterations. Even if an unrealistic value is sampled for one input, it may not produce an extreme value in the model output. If one does truncate a distribution, it can potentially affect the mean and other moments of the distribution. Thus, one expert summarized the issue of truncation as "nitpicking" that potentially can lead to more problems than it solves.

2.5 GOODNESS-OF-FIT

The fifth and final session of the workshop was devoted to the following trigger question:

On what basis should it be decided whether a data set is adequately fitted by a parametric distribution?

The premise of this session was the assumption that a decision had already been made to use a PDF instead of an EDF. While not all participating experts were comfortable with this assumption, all agreed to base the subsequent discussion on it.

The group agreed unanimously that visualization of both the data and the fitted distribution is the most important approach for ascertaining the adequacy of fit. The group in general seemed to share a view that conventional Goodness-of-Fit (GoF) tests have significant shortcomings and that they should not be the only or perhaps even primary methods for determining the adequacy of fit.

One expert elaborated that any type of probability plot that allows one to transform data so that they can be compared to a straight line, representing a perfect fit, is extremely useful. The human eye is generally good at identifying discrepancies from the straight line perfect fit. Another pointed out that visualization and visual inspection is routinely used in the medical community for evaluation of information such as x-rays and CAT scans; thus, there is a credible basis for reliance on visualization as a means for evaluating models and data.

One of the potential problems with GoF tests is that they may be sensitive to imperfections in the fit that are not of serious concern to an assessor or a decision maker. For example, if there are outliers at the low or middle portions of the distribution, a GoF test may suggest that a particular PDF should be rejected even though there is a good fit at the upper end of the distribution. In the absence of a visual inspection of the fit, the assessor may have no insight as to why a particular PDF was rejected by a GoF test.

The power of GoF tests was discussed. The group in general seemed comfortable with the notion of overriding the results of a GoF test if what appeared to be a good fit, via visual inspection, was rejected by the test, especially for large data sets or when the imperfections are in portions of the distribution that are not of major concern to the assessor or decision maker. Some experts shared stories of situations in which they found that a particular GoF test would reject a distribution due to only a few "strange" data points in what otherwise appears to be a plausible fit. It was noted that GoF tests become increasingly sensitive as the number of data points increases, so that even what appear to be small or negligible "blips" in a large data set are sufficient to lead to rejection of the fit. In contrast, for small data sets, GoF tests tend to be "weak" and may fail to reject a wide range of PDFs. One person expressed concern that any strict requirement for the use of GoF tests might reduce incentives for data collection, because it is relatively easy to avoid rejecting a PDF with few data.

The basis of GoF tests sparked some discussion. The "loss functions" assumed in many tests typically have to do with deviation of the fitted cumulative distribution function from the EDF for the data set. Other criteria are possible and, in principle, one could create any arbitrary GoF test. One expert asked whether minimization of the loss function used in any particular GoF test might be used as a basis for choosing parameter values when fitting a distribution to the data. There was no specific objection, but it was pointed out that a degree-of-freedom correction would be needed. Furthermore, other methods, such as maximum likelihood estimation (MLE), have a stronger theoretical basis as a method for parameter estimation.

The group discussed the role of the "significance level" and the "p-value" in GoF tests. One expert stressed that the significance level should be determined in advance of evaluating GoF and that it must be applied consistently in rejecting possible fits. Others, however, suggested that the appropriate significance level would depend upon risk management objectives. One expert suggested that it is useful to know the p-value of every fitted distribution so that one may have an indication of how good or weak the fit may have been according to the particular GoF test.

SECTION THREE

OPENING REMARKS

At the opening session of the workshop, representatives from EPA Region 2 and the RAF welcomed members of the expert panel and observers. Following EPA remarks, the workshop facilitator described the overall structure and objectives of the 2-day forum, which this section summarizes.

3.1 WELCOME AND REGIONAL PERSPECTIVE
Mr. William McCabe, Deputy Director, Program Support Branch, Emergency and Remedial Response Division, U.S. EPA Region 2

William McCabe welcomed the group to EPA Region 2 and thanked everyone for participating in the workshop. He noted that, in addition to this workshop, Region 2 also hosted the May 1996 Monte Carlo workshop, which ultimately led to the release of EPA's May 1997 policy document on probabilistic assessment. He commented on how this 2-day workshop was an important followup to the May 1996 event. Mr. McCabe stressed that continued discussions on viable approaches to probabilistic assessments are important because site-specific decisions rest on the merit of the risk assessment. He stated that this type of workshop is an excellent opportunity for attendees to discuss effective methods and expressed optimism that workshop discussions would provide additional insight and answers to probabilistic assessment issues. Resolution of key probabilistic assessment issues, he noted, will help the region members as they review risk assessments using probabilistic techniques. He mentioned, for example, the ongoing Hudson River PCB study for which deterministic and probabilistic assessments will be performed. In that case, as in others, Mr. McCabe said it will be critical for Agency reviewers to put the results into the proper context and to validate/critically review probabilistic techniques employed by the contractor(s) for the Potentially Responsible Parties.

3.2 OVERVIEW AND BACKGROUND
Mr. Steve Knott, U.S. EPA, Office of Research and Development, Risk Assessment Forum

On behalf of the RAF, Steve Knott thanked Region 2 for hosting the workshop. Mr. Knott briefly explained how the RAF originated in the early 1980s and comprises approximately 30 scientists from EPA program offices, laboratories, and regions. One primary RAF function is to bring experts together to carefully study and help foster cross-agency consensus on tough risk assessment issues.

Mr. Knott described the following activities related to probabilistic analysis in which the RAF has been involved:

- Formation of the 1983 ad hoc technical panel on Monte Carlo analysis.
- May 1996 workshop on Monte Carlo analysis (US EPA, 1996b).
- Development of the guiding principles for Monte Carlo analysis (US EPA, 1997a)
- EPA's general probabilistic analysis policy (US EPA, 1997b).

Mr. Knott reiterated the Agency's perspective on probabilistic techniques, stating that "the use of probabilistic techniques can be a viable statistical tool for analyzing variability and uncertainty in risk assessment" (US EPA, 1997b). Mr. Knott highlighted Condition 5 (on which this workshop was based) of the eight *conditions for acceptance* listed in EPA's policy:

> Information for each input and output distribution is to be provided in the report. This includes tabular and graphical representations of the distributions (e.g., probability density function and cumulative distribution function plots) that indicate the location of any point estimates of interest (e.g., mean, median, 95th percentile). *The selection of distributions is to be explained and justified.* For both the input and output distributions, variability and uncertainty are to be differentiated where possible (US EPA, 1997b).

Mr. Knott referred to the recent RTI report, "Development of Statistical Distributions for Exposure Factors" (1998), which presents a framework for fitting distributions and applies the framework to three case studies.

Mr. Knott explained that the Agency is seeking input from workshop participants primarily in the following areas:

- Methods for fitting distributions to less-than-perfect data (i.e., data that are not perfectly representative of the scenario(s) under study).
- Using the EDF (or resampling techniques) versus the PDF.

These issues were the focus of the workshop. Mr. Knott noted that the workshop will enable EPA to receive input from experts, build on existing guidance, and provide Agency assessors additional insight. EPA will use the information from this workshop in future activities, including (1) developing or revising guidelines and models, (2) updating the Exposure Factors Handbook, (3) supporting modeling efforts, and (4) applying probabilistic techniques to dose-response assessment.

3.3 WORKSHOP STRUCTURE AND OBJECTIVES
Dr. H. Christopher Frey, Workshop Chair

Dr. Frey, who served as workshop chair and facilitator, reiterated the purpose and goals of the workshop. As facilitator, Dr. Frey noted, he would attempt to foster discussions that would further illuminate and support probabilistic assessment activities. Dr. Frey stated that workshop discussions would center on the two issue papers mentioned previously. He explained that the RTI report was provided to experts for background purposes only. While the RTI report was not the review subject for this workshop, Dr. Frey commented that it may provide pertinent examples.

The group's charge, according to Dr. Frey, was to advise EPA and the profession on *representativeness* and *distribution function* issues. Because a slightly greater need exists for discussing representativeness issues and developing new techniques in this area, Dr. Frey explained that this topic would receive the greatest attention during the 2-day workshop. He reemphasized that the workshop would focus on technical issues, not policy issues.

Dr. Frey concluded his introductory remarks by stating that the overall goal of the workshop was to provide a framework for addressing technical issues that may be applied widely to different future activities (e.g., development of exposure factor distributions).

Workshop Structure and Expert Charge

Dr. Frey explained that the workshop would be structured around technical questions related to the two issue papers. Appendix D presents the charge provided to experts before the workshop, including specific questions for consideration and comment. The workshop material, Dr. Frey noted, is inherently technical. He, therefore, encouraged the experts to use plain language where possible. He also noted that the workshop was not intended to be a short course or tutorial. In introducing the key topics for workshop discussions, Dr. Frey highlighted the following, which he perceived as the most challenging issues and questions based on experts' premeeting comments:

> ***Representativeness.*** How should assessors address representativeness? What deviation is acceptable (given uncertainty and variability in data quality, how close will we come to answering the question)? How do assessors work representativeness into their problem definition (e.g., What are we asking? What form will the answer take?)
>
> ***Sensitivity.*** How important is the potential lack of representativeness? How do we evaluate this?
>
> ***Adjustment.*** Are there reasonable ways to adjust or extrapolate in cases where exposure data are not representative of the population of concern?
>
> ***EDF/PDF.*** How do assessors choose between EDFs and theoretical PDFs? On what basis do assessors decide whether a data set is adequately represented by a fitted analytic distribution?

Dr. Frey encouraged participants to remember the following general questions as they discussed specific technical questions during plenary sessions, small group discussions, and brainwriting sessions:

- What do we know today that we can apply to answer the questions or provide guidance?
- What short-term studies (e.g., numerical experiments) could answer the question or provide additional guidance?
- What long-term research (e.g., greater than 18 months) may be needed to answer the question or provide additional guidance?

According to Dr. Frey, the answers to these questions will help guide Agency activities related to probabilistic assessments.

Dr. Frey also encouraged the group to consider what, if anything, is not covered in the issue papers, but is related to the key topics. He noted some of the following examples, which were communicated in the experts' premeeting comments:

- Role of expert judgment and Bayesian methods, especially in making adjustments.

- Is model output considered representative if all the inputs to the model are considered representative? This issues relates, in part, to whether or not correlations or dependencies among the input are properly addressed.

- Role of representativeness in a default or generic assessment.

- Role of the measurement process.

Lastly, Dr. Frey explained that the activities related to the workshop are public information. The workshop was advertised in the Federal Register and observers were welcomed. Time was set aside on both days of the workshop for observer questions and comments.

SECTION FOUR

ISSUE PAPER PRESENTATIONS

Two issue papers were developed to present the expert panelists with pertinent issues and to initiate workshop discussions. Prior to the plenary and small group discussions, EPA provided an overview of each paper. This section provides a synopsis of each presentation. The two issue papers are presented in Appendix A. The overheads are in Appendix H.

4.1 ISSUE PAPER ON EVALUATING REPRESENTATIVENESS OF EXPOSURE FACTORS DATA

Jacqueline Moya, U.S. EPA, NCEA, Washington, DC

Ms. Moya opened her overview by noting that, while exposure distributions are available in the Exposure Factors Handbook, there is still a need to fit distributions for these data. Ms. Moya noted that a joint NCEA-RTI pilot project in September 1997 was established to do this. She then discussed the purpose of the issue paper and the main topics she planned to cover (i.e., framework for inferences, components of representativeness, the checklists, and methods for improving representativeness). The purpose of the issue paper, Ms. Moya reminded the group, was to introduce concepts and to prompt discussions on how to evaluate representativeness and what to do if a sample is *not* representative.

Ms. Moya presented a flow chart (see Figure 1 in the issue paper) of the data-collection process for a risk assessment. If data collection is not possible, she explained, surrogate data must be identified. The next step is to ask whether the surrogate data represent the site or chemical. Ms. Moya pointed to Checklist I (Assessing Internal Representativeness), which includes suggested questions for determining whether the surrogate data are representative of the population of concern. If not, the assessor must ask, "How do we adjust the data to make it more representative?"

Ms. Moya then briefly reviewed the key terms in the paper. *Representativeness* in the context of an exposure/risk assessment refers to the comfort with which one can draw inferences from the data. *Population* is defined in terms of its member characteristics (i.e., demographics, spatial and temporal elements, behavioral patterns). The assessor's *population of concern* is the population for which the assessment is being conducted. The *surrogate population* is the population used when data on the population of concern is not available. The *population of concern for the surrogate study* is the sample population for which the surrogate study was designed. The *population sampled* is a sample from the population of concern of the surrogate study.

Ms. Moya briefly described the external and internal components of representativeness. She explained that external components reflect how well the surrogate population represents the population of concern. Internal components refer to the surrogate study, specifically:

1. How well do sampled individuals represent the surrogate population? This depends on how well the study was designed. For example, was it random?

2. How well do the respondents represent the sample population? For example, if recreational fishermen are surveyed, is someone who fishes more frequently more likely to respond the survey, and therefore bias the response?

3. How well does the measured value represent the true value for the measurement unit? For example, are the recreational fishermen in the previous example accurately reporting the sizes of the fish they catch?

Ms. Moya reviewed the four checklists in the issue paper which may serve as tools for risk assessors trying to evaluate data representativeness. One checklist is for the population sampled versus the population of concern for the surrogate study (internal representativeness). The other checklists refer to the surrogate population versus the population of concern based on individual, spatial, and temporal characteristics (external representativeness). One goal of the workshop, Ms. Moya explained, was to solicit input from experts on the use of these checklists. Specifically, she asked whether certain questions should be eliminated (e.g., only a subset of the questions may be needed for a screening risk assessment).

Lastly, Ms. Moya pointed to discussions in the issue paper on attempting to improve representativeness. One section refers to how to make *adjustments* for differences in population characteristics (with discussions geared toward using weights for the sample). The second section refers to time-unit differences and includes how to adjust for this. Ms. Moya asked the group to consider how to evaluate the significance of population differences and how to perform extrapolations if they are necessary.

4.2 ISSUE PAPER ON EMPIRICAL DISTRIBUTION FUNCTIONS AND NON-PARAMETRIC SIMULATION
Timothy Barry, U.S. EPA, NCEA, Washington, DC

Dr. Barry reviewed the issues of concern related to selecting and evaluating distribution functions. He explained that, assuming data are representative, the risk assessor has two methods for representing an exposure factor in a probabilistic analysis: *parametric* (e.g., a Lognormal, Gamma, or Weibull distribution) and *non-parametric* (i.e., use the sample data to define an EDF).

To illustrate how the EDF is generated, Dr. Barry presented equations and histograms (see Appendix H). The basic EDF properties were defined as follows:

- Values between any two consecutive samples, x_k and x_{k+1}, cannot be simulated, nor can values smaller than the sample minimum, x_1, or larger than the sample maximum, x_n, be generated (i.e., $x>x_1$ and $x<x_n$).

- The mean of the EDF equals the sample mean. The variance of the EDF mean is always smaller than the variance of the sample mean; it equals *(n-1)/n* times the variance of the sample mean.

- Expected values of simulated EDF percentiles are equal to the sample percentiles.

- If the underlying distribution is skewed to the right (as are many environmental quantities), the EDF tends to underestimate the true mean and variance.

In addition to the basic EDF, Dr. Barry explained, the following variations exist:

- *Linearized EDF.* In this case, a linearized cumulative distribution pattern results. The linearized EDF linearly extrapolates between two observations.

- *Extended EDF.* An extended EDF involves linearization and adds lower and upper tails to the data to reflect a "more realistic range" of the exposure variable. Tails are added based expert judgment.

- *Mixed Exponential.* In this case, an exponential upper tail is added to the EDF. This approach is based on extreme value theory.

After describing the basic concepts of EDFs, Dr. Barry provided an example in which investigators compared and contrasted parametric and non-parametric techniques. Specifically, 90 air exchange data points were shown to have a Weibull fit. When a basic EDF for these data is used, means and variance reproduce well. It was concluded that if the goal is to reproduce the sample, Weibull does well on the mean but poorly at the high end.

Dr. Barry encouraged the group to consider the following questions during the 2-day workshop:

- Is an EDF preferred over a PDF in any circumstances?

- Should an EDF not be used in certain situations?

- When an EDF is used, should the linearized, extended, or mixed version be used?

Dr. Barry briefly described the Goodness of Fit (GoF) questions the issue paper introduces. He explained that, generally, assessors should pick the simplest analytic distribution not rejected by the data. Because rejection depends on the chosen statistic and on an arbitrary level of statistical significance, Dr. Barry posed the following questions to the group:

- What role should the GoF statistic and its p-value (when available) play in deciding on the appropriate distribution?

- What role should graphical assessments of fit play?

- When none of the standard distributions fit well, should you investigate more flexible families of distributions (e.g., four parameter gamma, four parameter F, mixtures)?

SECTION FIVE

EVALUATING REPRESENTATIVENESS OF EXPOSURE FACTORS DATA

Discussions on the first day and a half of the workshop focused on developing a framework for characterizing and evaluating the *representativeness* of exposure data. The framework described in the issue paper on representativeness (see Appendix A) is organized into three broad sets of questions: (1) those related to differences in populations, (2) those related to differences in spatial coverage and scale, and (3) those related to differences in temporal scale. Therefore, discussions were held in the context of these three topic areas. The panel also discussed the strengths and weaknesses of the proposed "checklists" in the issue paper, which were designed to help the assessor evaluate representativeness. The last portion of the workshop session on representativeness included discussions on sensitivity (assessing the importance of non-representativeness) and on the methods available to adjust data to better represent the population of concern. This section describes the outcome of each of these discussions.

Initial deliberations centered on the need to define risk assessment objectives (i.e. problem definition) before evaluating the representativeness of exposure data. Discussions on sensitivity and adjustment followed.

5.1 PROBLEM DEFINITION

The group agreed on two points: that "representativeness" depends on the problem at hand and that the context of the risk analysis is critical. Several experts commented that assessors will have a difficult time defining representativeness if the problem has not been well-defined. The group therefore spent a significant amount of time discussing problem definition and problem formulation in the context of assessing representativeness. Several experts noted the importance of understanding the *end use* of the assessment (e.g., site-specific or generic, national or regional analysis). The group agreed that the most important step for assessors is to ask whether the data are representative enough for their intended use(s).

The group agreed that stakeholders and other data users should be involved in all phases of the assessment process, including early brainstorming sessions. Two experts noted that problem definition must address whether the assessment will adequately protect public health and the environment. Another expert stressed the importance of problem formulation, because not doing so risks running analyses or engaging resources needlessly. One participant commented that the importance of representativeness varies with the level (or tier) of the assessment. For example, if data are to be used in a screening manner, then conservativeness may be more important than representativeness. If data are to be used in something other than screening assessments, the assessor must consider *the value added* of more complex analyses (i.e., additional site-specific data collection, modeling). Two experts noted, however, that the following general problem statement/question would not change with a more or less sophisticated (tiered) assessment: Under an agreed upon set of exposure conditions, will the population of concern experience unacceptable risks? A more sophisticated analysis would merely enable a closer look at less conservative/more realistic conditions.

5.1.1 What information is required to specify a problem definition fully?

The group agreed that when defining any problem, the "fundamental who, what, when, where, why, and how" questions must be answered. One individual noted that if assessors answer these questions, they will be closer to determining if data are representative. The degree to which each basic question is important is specific to the problem or situation. Another reiterated the importance of remembering that the premier consideration is public health protection; he noted that if only narrow issues are discussed, the public health impact may be overlooked.

The group concurred that the problem must be defined in terms of location (space), time (over what duration and when in time), and population (person or unit). Some of these definitions may be concrete (e.g., spatial locations around a site), while some, like people who live on a brownfield site, may be more vague (e.g., because they may change with mobility and new land use). Because the problem addresses a future context, it must be linked to observable data by a model and assumptions. The problem definition should include these models and assumptions.

Various experts provided the following specific examples of the questions assessors should consider at the problem formulation stage of a risk assessment.

- What is the purpose of the assessment (e.g., regulatory decision, setting cleanup standards)?
- What is the population of interest?
- What type of assessment is being performed (site-specific or generic)?
- How is the assessment information being used? How will data be used (e.g., screening assessment versus court room)?
- Who are the stakeholders?
- What are the budget limitations? What is the cost/benefit of performing a probabilistic versus a deterministic assessment?
- What population is exposed, and what are its characteristics?
- How, when, and where are people exposed?
- In what activities does the exposed population engage? When does the exposed population engage in these activities, and for how long? Why are certain activities performed?
- What type of exposure is being evaluated (e.g., chronic/acute)?
- What is the scenario of interest (e.g., what is future land use)?
- What is the target or "acceptable" level of risk (e.g., 10^{-2} versus 10^{-6})?

- What is the measurement error?
- What is the acceptable level of error?
- What is the geographic scale and location (e.g., city, county)?
- What is the scale for data collection (e.g., regional/city, national)?
- What are site/region-specific issues (e.g., how might a warm climate or poor-tasting water affect drinking water consumption rates)?
- What is the temporal scale (day, year, lifetime)?
- What are the temporal characteristics of source emissions (continuous)?
- What is/are the route(s) of exposure?
- What is the dose (external, biological)?
- What is/are the statistic(s) of interest (e.g., mean, uncertainty percentile)?
- What is the plausible worst case?
- What is the overall data quality?
- What models must be used?
- What is the measurement error?
- When would results change a decision?

Many of the preceding questions are linked closely to defining representativeness. One subgroup compiled a list of key elements that are directly related to these types of questions when defining representativeness (see textbox on page 5-4).

5.1.2 What constitutes representativeness (or lack thereof)? What is "acceptable deviation"?

Several of the experts commented that, fundamentally, representativeness is a function of the quality of the data but reiterated that it depends ultimately on the overall assessment objective. Almost all data used in risk assessment fail to be representative in one or more ways. At issue is the effect of the lack of representativeness on the risk assessment. One expert suggested that applying the established concepts of EPA's data quality objective/data quality assessment process would help assessors evaluate data representativeness. Because populations are not fixed in time, one expert cautioned that if a data set is too representative, the risk assessment may be precise for only a moment. Another stressed the importance of taking a credible story to the risk manager. In that context, "precise representativeness" may be less important than answering the question of whether we are being protective of public health. It

Sources of Variability and Uncertainty Related to the Assessment of Data Representativeness

EPA policy sets the standard that risk assessors should seek to characterize central tendency and plausible upper bounds on both individual risk and population risk for the overall target population as well as for sensitive subpopulations. To this extent, data representativeness cannot be separated from the assessment endpoint(s). Following are some key elements that may affect data representativeness. These elements are not mutually exclusive.

Exposed Population

- General target population
- Particular ethnic group
- Known sensitive subgroup (e.g., children, elderly, asthmatics)
- Occupational group (e.g., applicators)
- Age group (e.g., infant, child, teen, adult, whole life)
- Gender
- Activity group (e.g., sport fishermen, subsistence fishermen)

Geographic Scale, Location

- Trends (e.g., stationary, nonstationary behaviors)
- Past, present, future exposures
- Lifetime exposures
- Less-than-lifetime exposures (e.g., hourly, daily, weekly, annually)
- Temporal characteristics of source(s) (e.g., continuous, intermittent, periodic, concentrated, random)

Exposure Route

- Inhalation
- Ingestion (e.g., direct, indirect)
- Dermal (direct) contact (by activity; e.g., swimming)
- Multiple pathways

Exposure/Risk Assessment Endpoint

- Cancer risk
- Noncancer risk (margin of exposure, hazard index)
- Potential dose, applied dose, internal dose, biologically effective dose
- Risk statistic
- Mean, uncertainty percentile of mean
- Percentile of a distribution (e.g., 95th percentile risk)
- Uncertainty limit of variability percentile (upper confidence limit on 95th percentile risk)
- Plausible worst case, uncertainty percentile of plausible worst case

Data Quality Issues

- Direct measurement, indirect measurement (surrogates)
- Modeling uncertainties
- Measurement error (accuracy, precision, bias)
- Sampling error (sample size, non-randomness, independence)
- Monitoring issues (short-term, long-term, stationary, mobile)

is important to understand whether a lack of representativeness could mean the risk assessment results fail to protect public health or that they grossly overestimate risks.

One participant expressed concern that assessors feel deviations from representativeness can be measured. In reality, risk assessors may more often rely on qualitative or semiquantitative ways of describing that deviation. Another expert emphasized that assessors often have no basis on which to judge the representativeness of surrogate data (e.g., drinking water consumption), because rarely is local data available for comparison. Therefore, surrogate data, must be accepted or modified based on some qualitative information (e.g., the local area is hotter than that which the surrogate data is based).

The experts provided the following views on what constitutes representativeness and/or an acceptable level of non-representativeness. These views were communicated during small group and plenary discussions.

Nearly consistent with the definition in the issue paper, *representativeness* was defined by one subgroup as "the degree to which a value for a given endpoint *adequately* describes the value of that endpoint(s) likely seen in the target population." The term "adequately" replaces the terms "accurately and precisely" in the issue paper definition. One expert suggested changing the word representative to "useful and informative." The latter terms imply that one has learned something from the surrogate population. For example, the assessor may not prove the data are the same, but can, at minimum, capture the extent to which they differ. The term *non-representativeness* was defined as *"important differences* between target and surrogate populations with respect to the risk assessment objectives." Like others, this subgroup noted that the context of observation is important (e.g., what is being measured: environmental sample [water, air, soil] versus human recall [diet] versus tissue samples in humans [e.g., blood]). Assessors must ask about internal sample consistency, inappropriate methods, lack of descriptors (e.g., demographic, temporal), and inadequate sample size for targeted measure.

The group agreed, overall, that assessing adequacy or representativeness is inherently subjective. However, differing opinions were offered in terms of how to address this subjectivity. Several participants stressed the importance of removing subjectivity to the extent possible but without making future guidance too rigid. Others noted, however, that expert judgment is and must remain an integral part of the assessment process.

A common theme communicated by the experts was that representativeness depends on how much uncertainty and variability between the population of concern and the surrogate population the assessor is willing to accept. What is "good enough" is case specific, as is the "allowable error." Several experts commented that it is also important for assessors to know if they are comparing data means or tails. One expert suggested reviewing some case studies using assessments done for different purposes to illuminate the process of defining representativeness. "With regard to exposure factors, we [EPA] need to do a better job at specifying or providing better guidance on how to use the data that are available." For example, the soil ingestion data for children are limited, but they may suffice to provide an estimate of a mean. These data are not good enough to support a distribution or a good estimate of a high-end value, however.

One subgroup described representativeness/non-representativeness as the degree of bias between a data set and the problem. For example:

Scenario: Is a future residential scenario appropriate to the problem? For prospective risk assessment, there are usually irreducible uncertainties about making estimates about a future unknown population. Therefore, a certain amount of modeling must occur.

Model: Is a multiplicative, independent variable model appropriate? Uncertainties in the model can contribute to non-representativeness (e.g., it might not apply, it may be wrong, or calculations may be incorrect).

Variables: Is a particular study appropriate to the problem at hand—are the variables biased, uncertain? It may be easy to get confused about distinctions between bias (or inaccuracies), precision/imprecision, and representativeness/non representativeness. It is often assumed that a "representative" data set is one that has been obtained with a certain amount of randomization. More often, however, data that meet this definition are not available.

The group spokesperson explained that a well-designed and controlled randomized study yielding two results can be "representative" of the mean and dispersion but highly imprecise. Imprecision and representativeness are therefore different, but related. The central tendency of the distribution may be accurately estimated, but the upper percentile may not.

In summary, when assessing representativeness, the group agreed that emphasis should be placed on the *adequacy* of the data and how *useful and informative* a data set is to the defined problem. The group agreed that these terms are more appropriate than "accuracy and precision" in defining representative data in the context of a risk assessment. The importance of considering end use of the data was stressed and was a recurring theme in the discussions (i.e., how much representativeness is needed to answer the problem). Because the subject population is often a moving target with unpredictable direction in terms of its demographics and conditions of exposure, one expert commented that, in some cases, representativeness of a given data set may not be a relevant concept and generic models may be more appropriate.

5.1.3 What considerations should be included in, added to, or excluded from the checklists?

More than half the experts indicated that the checklists in Issue Paper 1 are useful for evaluating representativeness. One expert noted that regulators are often forced to make decisions without information. A checklist helps the assessor/risk manager evaluate the potential importance of missing exposure data. One expert re-emphasized the importance of allowing for professional judgement and expert elicitation when evaluating exposure data. Another panelist concurred, commenting that this type of the checklist is preferred over prescriptive guidance. Several of the experts noted, however, that checklists could be improved and offered several recommendations.

The group agreed that the checklist should be flexible for various problems and that users should be directed to consider the purpose of the risk assessment. The assessor must know the minimum requirements for a screening versus a probabilistic assessment. As one expert said, the requirements for a screening level assessment must differ from those for a full-blown risk assessment: Do I have enough information about the population (e.g., type, space, time) to answer the questions at this tier, and is that information complete enough to make a management decision? Do I need to go through all the checklists before I can stop?

Instead of the binary (yes/no) and linear format of the checklists, several individuals suggested a flowchart format centered on the critical elements of representativeness (i.e., a "conditional" checklist)—to what extent does the representativeness of the data really matter? A flowchart would allow for a more iterative process and would help the assessor work through problem-definition issues. One expert suggested developing an interactive Web-based flowchart that would be flexible and context-specific. Another agreed, adding that criteria are needed to guide the assessor on what to do if information is not available. As one expert noted, questions should focus on the outcome of the risk assessment. The assessor needs to evaluate whether the outcome of the assessment changes if the populations differ.

One of the experts strongly encouraged collecting more/new data or information. Collection of additional data, he noted, is needed to improve the utility of these checklists. Another participant suggested that the user be alerted to the qualities of data that enable quantifying uncertainty and reminded that the degree of representativeness cannot be defined in certain cases. When biases due to lack of representativeness are suspected, how can assessors judge the direction of those biases?

In addition to general comments and recommendations, several individuals offered the following specific suggestions for the checklists:

- Clarifying definitions (e.g., internal versus external).
- Recategorizing. For example, use the following five categories: (1) interpreting measurements (more of a validity than representative issue), (2) evaluating whether sampling bias exists, (3) evaluating statistical sampling error, (4) evaluating whether the study measured what must be known, and (5) evaluating differences in the population. The first three issues are sources of internal error, the latter two are sources of external representativeness.
- Reducing the checklists. Several experts suggested combining Checklists II, III, and IV.
- Combining temporal, spatial, and individual categories. Avoid overlap in questions. For example, when overlap exists (e.g., in some spatial and temporal characteristics), which questions in the checklist are critical? A Web-based checklist, with the flow of questions appropriately programmed, could be designed to avoid duplication of questions.
- Including other populations of concern (e.g., ecological receptors).
- Including worked examples that demonstrate the criteria for determining if a question is answered adequately and appropriately. These examples should help focus the risk assessor on the issues that are critical to representativeness.
- Separating bias and sampling quality and extrapolation from reanalysis and reinterpretation.

- Asking the following additional questions:
  - Relative to application, is there consistency in the survey instruments used to collect the exposure data? How was measurement error addressed?
  - Is the sample representative enough to bound the risk?
  - Are data available on population characterization factors (e.g., age, sex)?
  - What is known about the population of concern relative to the surrogate population? (If the population of concern is inadequately characterized, then the ability to consider the representativeness of the surrogate data is limited, and meaningless adjustment may result).

In summary, the group agreed on the utility of the checklists but emphasized the need to include in them decision criteria (i.e., how do we know if we have representative/non-representative data?) A brief discussion on the need to collect data followed. Some experts posed the following questions: How important is it to have more data? Is the risk assessment really driving decisions? Is more information needed to make good decisions? Is making risk assessment decisions on qualitative data acceptable? What data must to be collected, at minimum, to validate key assumptions? The results of the sensitivity analysis, as one expert pointed out, are key to answering these questions.

5.2 SENSITIVITY

How do we assess the importance of non-representativeness?

In considering the implications of non-representativeness, the group was asked to consider how one identifies the implications of non-representativeness in the context of the risk assessment. One expert commented that the term "non-representativeness" may be a little misleading, and as discussed earlier, finds the terms *data adequacy* or *data useability* more fitting to the discussions at hand. The expert noted that, from a Superfund perspective, data representativeness is only one consideration when assessing overall data quality or useability. Others agreed. The workshop chair encouraged everyone to discuss the suitability of the term "representativeness" while assessing its importance during the small group discussions.

One group described a way in which to assess the issue of non-representativeness as follows: The assessor must check the sensitivity of decisions to be made as a result of the assessment. That is, under a range of plausible adjustments, will the risk decision change? Representativeness is often not that important because risk management decisions depend on a range of target populations under various scenarios. A few of the experts expressed concern that problems will likely arise if the exposure assessor is separated from decision makers. One person noted that often times an exposure assessment will be done absent of a specific decision (e.g., nonsite, non-Superfund situation). Another noted that in the pesticide program situations occur in which an exposure assessment is done before toxicity data are available. Such separations may be unavoidable. Another expert emphasized that any future guidance should stress the importance of assessors being cognizant of data distribution needs even if the assessors are removed from the decision or have limited data.

One individual noted that examples would help. The assessor should perform context-specific sensitivity analysis. It would help to develop case studies and see how sensitivity analysis affects application (e.g., decision focus).

Another group discussed sensitivity analysis in the context of a tiered approach. For the first tier, a value that is "biased high" should be selected (e.g., 95th percentile upper bound). The importance of a parameter (as evidenced by a sensitivity analysis) is determined first, making the representativeness or non-representativeness of the nonsensitive parameters unimportant. For the second tier (for sensitive parameters), the assessor must consider whether averages or high end estimates are of greater importance. This group presented an example using a corn oil scenario to illustrate when differences between individuals (e.g. high end) and mixtures (averages) may be important. Because corn oil is a blend with input from many ears of corn, if variability exists in the contaminant concentrations in individual ears of corn, then corn oil will typically represent some type of average of those concentrations. For such a mixture, representativeness is less of an issue. It is not necessary to worry about peak concentrations in one ear of corn. Instead, one would be interested in situations which might give rise to a relatively high average among the many ears of corn that comprise a given quantity of corn oil. If one is considering individual ears of corn, it becomes more important to have a representative sample; the tail of the distribution becomes of greater interest.

A third subgroup noted that, given a model and parameters, assessors must determine whether enough data exist to bound the estimates. If they can bound the estimates, a sensitivity analysis is performed with the following considerations: (1) identify the sensitive parameters in the model; (2) focus on sensitive parameters and evaluate the distribution beyond the bounding estimate (i.e., identify the variability of these parameters) for the identified sensitive parameters; (3) evaluate whether the distribution is representative; and (4) evaluate whether more data should be collected or if an adjustment is appropriate.

Members of the remaining subgroup noted, and others agreed, that a "perfect" risk assessment is not possible. They reiterated that it is key to evaluate the data in the context of the decision analysis. Again, what are the consequences of being wrong, and what difference do decision errors make in the estimate of the parameter being evaluated? This group emphasized that the question is situation-specific. In addition, they noted the need for placing bounds on data used.

One question asked throughout these discussions was "Are the data good enough to replace an existing assumption and, if not, can we obtain such data?" One individual again stressed the need for "blue chip" distributions at the national level (e.g., inhalation rate, drinking water). Another expert suggested adding activity patterns to the list of needed data.

In summary, the group generally agreed that the sensitivity of the risk assessment decision must be considered before non-representativeness is considered problematic. In some cases, there may not be an immediate decision, but good distributions are still important.

How can one do sensitivity analysis to evaluate the implications of non-representativeness?

The workshop chair asked the group to consider the mechanics of a sensitivity analysis. For example, is there a specific statistic that should be used, or is it decision dependent? One expert responded by noting that sensitivity analysis can be equated to partial correlation coefficients (which are internal to a model). He noted, however, that sensitivity analysis in the context of exposure assessment is more "bottom line" sensitivity (i.e., if an assumption is changed, how does the change affect the bottom line?). The focus here is more external—what happens when you change the inputs to the model (e.g., the distributions)? Another pointed to ways in which to perform internal sensitivity analysis. For example, the sensitivity of uncertainty can be separated out from the sensitivity of the variability component (see William Huber's premeeting comments on sensitivity). Another expert stressed, however, that sensitivity analysis is inherently tied to uncertainty; it is not tied to variability unless the variability is uncertain. It was noted that sensitivity analysis is an opportunity to view things that are subjective. Variability, in contrast is inherent in the data, unless there are too few data to estimate variability sufficiently. One expert commented that it is useful to know which sources of variability are most important in determining exposure and risk.

One individual voiced concern regarding how available models address sensitivity. Another questioned whether current software (e.g., Crystal Ball® and @Risk®) covers sensitivity coefficients adequately (i.e., does it reflect the depth and breadth of existing literature?).

Lastly, the group discussed sensitivity analysis in the context of what we know now and what we need to know to improve the existing methodology. Individuals suggested the following:

- Add the ability to classify sample runs to available software. Classify inputs and evaluate the effect on outputs.

- Crystal Ball® and @Risk® are reliable for many calculations, but one expert noted they may not currently be useful for second-order estimates, nor can they use time runs. Time series analyses are particularly important for Food Quality Protection Act (FQPA) evaluations.

- Consider possible biases built into the model due to residuals lost during regression analyses. This factor is important to the sensitivity of the model prediction.

One expert pointed out that regression analyses can introduce bias because residuals are often dropped out. Others agreed that this is an important issue. For example, it can make an order-of-magnitude difference in body weight and surface area scaling. Another expert stated that this issue is of special interest for work under the FQPA, where use of surrogate data and regression analysis is receiving more and more attention. Another expert noted that "g-estimation" looks at this issue. The group revisited this issue during their discussions on adjustment.

5.3 ADJUSTMENT

How can one adjust the sample to better represent the population of interest?

The experts addressed adjustment in terms of population, spatial, and temporal characteristics. The group was asked to identify currently available methods and information sources that enable the quantitative adjustment of surrogate sample data. In addition, the group was asked to identify both short- and long-term research needs in this area. The workshop chair noted that the issue paper only includes discussion on adjustments to account for time-scale differences. The goal, therefore, was to generate some discussion on spatial and population adjustments as well. Various approaches for making adjustments were discussed, including general and mechanistic. General approaches include those that are statistically-, mathematically-, or empirically-based (e.g., regression analysis). Mechanistic approaches would involve applying a theory specific to a problem area (e.g., a biological, chemical, or physical model).

Some differing opinions were provided as to how reliably we can apply available statistics to adjust data. In time-space modeling, where primary data and multiple observations occur at different spatial locations or in multiple measures over time, one expert noted that a fairly well-developed set of analytic methods exist. These methods would fall under the category of mix models, kreiging studies for spatial analysis, or random-effects models. The group agreed that extrapolating or postulating models are less well-developed. One person noted that classical statistics fall short because they do not apply to situations in which representativeness is a core concern. Instead, these methods focus more on the accuracy or applicability of the model. The group agreed that statistical literature in this area is growing.

Another individual expressed concern that statistical tools and extrapolations introduce more uncertainty to the assessment. This uncertainty may not be a problem if the assessor has good information about the population of concern and is simply adjusting or reweighing the data, but when the assessor is extrapolating the source term, demographics, and spatial characteristics simultaneously, more assumptions and increasing uncertainty are introduced.

In general, the group agreed that a model-based approach has merit in certain cases. The modeled approach, as one expert noted, is a cheap and effective approach and likely to support informed/more objective decisions. The group agreed that validated models (e.g., spatial/fate and transport models) should be used. Because information on populations may simply be unavailable to validate some potentially useful models, several participants reemphasized the need to collect more data, which was a recurring workshop theme.

One expert pointed out that the assessor must ask which unit of observation is of concern. For example, when evaluating cancer risk, temporal/spatial issues (e.g., residence time) are less important. When evaluating developmental effects (when windows of time are important), however, the temporal/spatial issues are more relevant. Again, assessors must consider the problem at hand before identifying the unit of time.

From a pesticide perspective, it was noted that new data cannot always be required of registrants. When considering the effects of pesticides, for example, crop treatment rates change over time. As a result, bridging studies are used to link available application data to crop residues (using a multiple linear regression model).

One expert stressed the importance and need for assessors to *recognize uncertainty*. Practitioners of probabilistic assessment should be encouraged to aggressively evaluate and discuss the uncertainties in extrapolations and their consequences. Often, probabilistic techniques can provide better information for better management decisions. The expert pointed out that, in some cases, one may not be able to assign a distribution, or one may choose not to do so because it would risk losing valuable information. In those cases, multiple scenarios and results reported in a nonprobabilistic way (both for communication and management decisions) may be appropriate.

At this point, one expert suggested that the discussion of multiple scenarios was straying from the basic question to be answered— "If I have a data set that does not apply to my population, what do I need to do, if anything?" Others disagreed, noting that it may make sense to run different scenarios and evaluate the difference. If a different scenario makes a difference, more data must be collected. One expert argued, however, that we cannot wait to observe trends; assessors must predict the future based on a "snapshot" of today.

One expert suggested the following hierarchy when deciding on the need to refine/adjust data:

- Can the effect be bounded? If yes, no adjustment is needed.
- If the bias is conservative, no adjustment is needed.
- Use a simple model to adjust the data.
- If adjustments fail, resample/collect more data, if possible.

The group then discussed the approaches and methods that are currently available to address non-representative data, and indicated that the following approaches are viable:

1. Start with brainstorming. Obtain stakeholder input to determine how the target population differs from the population for which you have data.

2 Look at covariates to get an idea of what adjustment might be needed. Stratify data to see if correlation exists. Stratification is a good basis for adjustments.

3. Use "kreiging" techniques (deriving information from one sample to a smaller, sparser data set). Kreiging may not fully apply to spatial, temporal, and population adjustments, however, because it applies to the theory of random fields. Kreiging may help improve the accuracy of existing data, but it does not enable extrapolation.

4. Include time-steps in models to evaluate temporal trends.

5. Use the "plausible extrapolation" model. This model is acceptable if biased conservatively.

6. Consider spatial estimates of covariate data (random fields).

7 Use the scenario approach instead of a probabilistic approach.

8. Bayesian statistical methods may be applicable and relevant.

   One expert presented a brief case study as an example of Bayesian analysis of variability and uncertainty and use of a covariate probability distribution model based on regression to allow extrapolation to different target populations. The paper he summarized, "Bayesian Analysis of Variability and Uncertainty on Arsenic Concentrations in U.S. Public Water Supplies," and supporting overheads, are in Appendix G. The paper describes a Bayesian methodology for estimating the distribution and its dependence on covariates. Posterior distributions were computed using Markov Chain Monte Carlo (MCMC). In this example, uncertainties and variability were associated with time issues and the self-selected nature of arsenic samples. After briefly reviewing model specifications and distributional assumptions, the results and interpretations were presented, including a presentation of MCMC output plots and the posterior cumulative distribution of source water. The uncertainty of fitting site-specific data to the national distribution of arsenic concentrations was then discussed. The results suggest that Bayesian methodology powerfully characterizes variability and uncertainty in exposure factors. The probability distribution model with covariates provides insights and a basis for extrapolation to other targeted populations or subpopulations. One of the main points of presenting this methodology was to demonstrate the use of covariates. This case study showed that you can fit a model with covariates, explicitly account for residuals (which may be important), and apply that same model to a separate subpopulation where you know something about the covariates. According to the presenter, such an approach helps reveal whether national data represent local data.

When evaluating research needs, one expert pointed out that assessors should identify the minimal amount of information they need to analyze the data using available tools. The group offered the following suggestions for both short and long-term research areas. The discussion of short-term needs also included recommendations for actions the assessors can take now or in the short term to address the topics discussed in this workshop.

Short-term research areas and actions

1. Design studies for data collection that are amenable to available methods for data analysis. Some existing methods are unusable because not all available data, which were used to support the methods, are from well-designed studies.

2. Validate existing models on population variability (e.g., the Duan-Wallace model [Wallace et al., 1994] and models described by Buck et al. [1995]). This validation can be achieved by collecting additional data.

3. Run numerical experiments to test existing and new methods for making adjustment based on factors such as averaging times or area. Explore and evaluate the Duan-Wallace model.

4. Hold a separate workshop on adjustment methods (e.g., geostatistical and time series methods). Involve the modelers working with these techniques on a cross-disciplinary panel to learn how particular techniques might apply to adjustment issues that are specific to risk assessment.

5. Provide guidelines on how to evaluate or choose an available method, instead of simply describing available techniques. These guidelines would help the assessor determine whether a method applies to a specific problem.

6. To facilitate their access and utility, place national data on the Web (e.g., 3-day CSFII data, 1994–1996 USDA food consumption data). Ideally, the complete data set, not just summary data, could be placed on the Web because data in summary form is difficult to analyze (e.g., to fit distributions).

Possible long-term research areas

1. Collect additional exposure parameter data on the national and regional levels (e.g., "blue chip" distributions). One expert cautioned that some sampling data have been or may be collected by field investigators working independently of risk assessment efforts. Therefore, risk assessors should have input in methods for designing data collection.

2. Perform targeted studies (spatial/temporal characteristics) to update existing data.

Discussions of *adjustment* ended with emphasis on the fact that adjustment and the previously described methods *only need be considered if they impact the endpoint.* One expert reiterated that when no quantitative or objective ways exist to adjust the surrogate data, a more generalized screening approach should be used.

As a follow-up to the adjustment discussions, a few individuals briefly discussed the issue of "bias/loss function" to society. Because this issue is largely a policy issue, it only received brief attention. One expert noted that overconservatism is undesirable. Another stressed that it is not in the public interest to extrapolate in the direction of not protecting public health; assessors should apply conservative bias but make risk managers aware of the biases. The other expert countered that blindly applying conservative assumptions could result in suboptimal decisions, which should not be taken lightly. In general, the group agreed on the following point: Assessors should use their *best scientific judgment* and strive for accuracy when considering representativeness and uncertainty issues. Which choice will ensure protection of public health without unreasonable loss? It was noted that the cost of overconservatism should drive the data-collection push (e.g., encourage industry to contribute to data collection efforts because they ultimately pay for conservative risk assessments).

5.4 SUMMARY OF EXPERT INPUT ON EVALUATING REPRESENTATIVENESS

Workshop discussions on representativeness revealed some common themes. The group generally agreed that representativeness is context-specific. Methods must be developed to ensure representativeness exists in cases where lack of representativeness would substantially impact a risk-management decision. Methods, the sensitivity analysis, and the decision endpoint are closely linked. One expert warned that once the problem is defined, the assessor must understand how to use statistical tools properly to meet assessment goals. Blind application of these tools can result in wrong answers (e.g., examining the tail versus the entire curve).

One or more experts raised the following issues related to evaluating the quality and "representativeness" of exposure factors data:

- Representativeness might be better termed "adequacy" or "usefulness."

- Before evaluating representativeness, the risk assessor, with input from stakeholders, must define the assessment problem clearly.

- No data are perfect; assessors must recognize this fact, clearly present it in their assessments, and adjust non-representative data as necessary using available tools. The assessors must make plausible adjustments if non-representativeness matters to the endpoint.

- To perform a probabilistic assessment well, adequate data are necessary, even for an assessment with a well-defined objective. In large part, current exposure distribution data fall short of the risk assessors' needs. Barriers to collecting new data must be identified, then removed. Cost limitations were pointed out, however. One expert, therefore, recommended that justification and priorities be established.

- Methods must be sensitive to needs broader than the Superfund/RCRA programs (e.g., food quality and pesticide programs).

- When evaluating the importance of representativeness and/or adjusting for non-representativeness, the assessor needs to make decisions that are adequately protective of public health while still considering costs and other loss functions. Ultimately, the assessor should strive for accuracy.

Options for the assessor when the population of concern has been shown to have different habits than the surrogate population were summarized as follows: (1) determine that the data are clearly not representative and cannot be used; (2) use the surrogate data and clearly state the uncertainties; or (3) adjust the data, using what information is available to enable a reasonable adjustment.

[illegible]

[illegible] representativeness, the assessor needs to make decisions that are adequately [illegible] public health while still considering costs and other loss functions. Ultimately, the assessor should strive for accuracy.

Options for the assessor when the population of concern has been shown to have different habits than the surrogate population were summarized as follows: (1) determine that the data are clearly not representative and cannot be used; (2) use the surrogate data and clearly state the uncertainties; or (3) adjust the data, using what information is available to enable a reasonable adjustment

SECTION SIX

EMPIRICAL DISTRIBUTION FUNCTIONS AND RESAMPLING VERSUS PARAMETRIC DISTRIBUTIONS

Assessors often must understand and judge the use of parametric methods (e.g., using such theoretical distribution functions as the Lognormal, Gamma, or Weibull distribution) versus non-parametric methods (using an EDF) for a given assessment. The final session of the workshop was therefore dedicated to exploring the strengths and weaknesses of EDFs and issues related to judging the quality of fit for theoretical distributions. Discussions centered largely on the topics in Issue Paper 2 (see Appendix A for a copy of the paper and Section 3 for the workshop presentation of the paper). This section presents a summary of expert input on these topics.

Some of the experts thought the issue paper imposed certain constraints on discussions because it assumed that: (1) no theoretical premise exists for assuming a parametric distribution, and (2) the data are representative of the exposure factor in question (i.e., obtained as a simple random sample and in the proper scale). These experts noted that many of the assertions in the issue paper do not exist in reality. For example, it is unlikely to find a perfectly random sample for exposure parameter data.

As a result, the discussions that followed covered the relative advantages and disadvantages of parametric and non-parametric distributions under a broader range of conditions.

6.1 SELECTING AN EDF OR PDF

Experts were asked to consider the following questions.

What are the primary considerations in choosing between the use of EDFs and theoretical PDFs? What are the advantages of one versus the other? Is the choice a matter of preference? Are there situations in which one method is preferred over the other? Are there cases in which neither method should be used?

The group agreed that selecting an EDF versus a PDF is often a matter of personal preference or professional judgment. It is not a matter of systematically selecting either a PDF- or EDF-based approach for every input. It was emphasized that selection of a distribution type is case- or situation-specific. In some cases, both approaches might be used in a single assessment. The decision, as one expert pointed out, is driven largely by data-rich versus data-poor situations. The decision is based also on the risk assessment objective. Several experts noted that the EDF and PDF have different strengths in different situations and encouraged the Agency not to recommend the use of one over the other or to develop guidance that is too rigid. Some experts disputed the extent to which a consistent approach should be encouraged. While it was recognized that a consistent approach may benefit decision making, the overall consensus was that too many constraints would inhibit the implementation of new/innovative approaches, from which we could learn.

Technical discussions started with the group distinguishing between "bootstrap" methods and EDFs. One expert questioned if the methods were synonymous. EDF, as one expert explained, is a specific type of step-wise distribution that can be used as a basis for bootstrap simulations. EDF is one way to describe a distribution using data; bootstrapping enables assessors to resample that distribution in a special way (e.g., setting boundaries on the distribution of the mean or percentile) (Efron and Tibshirani, 1993). Another expert distinguished between a parametric and non-parametric bootstrap, stating that there are good reasons for using both methods. These reasons are well-covered in the statistical literature. One expert noted that bootstrapping enables a better evaluation of the uncertainty of the distribution.

Subsequent discussion focused on expert input on deciding which distribution to fit, if any, for a given risk assessment problem. That is, if the assessor has a data set that must be represented, is it better to use the data set as is and not make any assumptions or to fit the data set to a parametric distribution? The following is a compilation of expert input.

- *Use of the EDF.* The use of an EDF may be preferable (1) when a large number of data points exists, (2) when access is available to computers with high speed and storage capabilities, (3) when no theoretical basis for selecting a PDF exists, or (4) when a "perfect" data set is available. With small data sets, it was noted that the EDF is unlikely to represent an upper percentile adequately; EDFs are restricted to the range of observed data. One expert stated that while choice of distribution largely depends on sample size, in most cases he would prefer the EDF.

  When measurement or response error exists, one expert pointed out that an EDF should not be used before looking at other options.

- *Use of the PDF.* One expert noted that it is easier to summarize a large data set with a PDF as long as the fit is reasonable. Use of PDFs can provide estimates of "tails" of the distribution beyond the range of observed data. A parametric distribution is a convenient way to concisely summarize a data set. That is, instead of reporting the individual data values, one can report the distribution and estimated parameter values of the distribution.

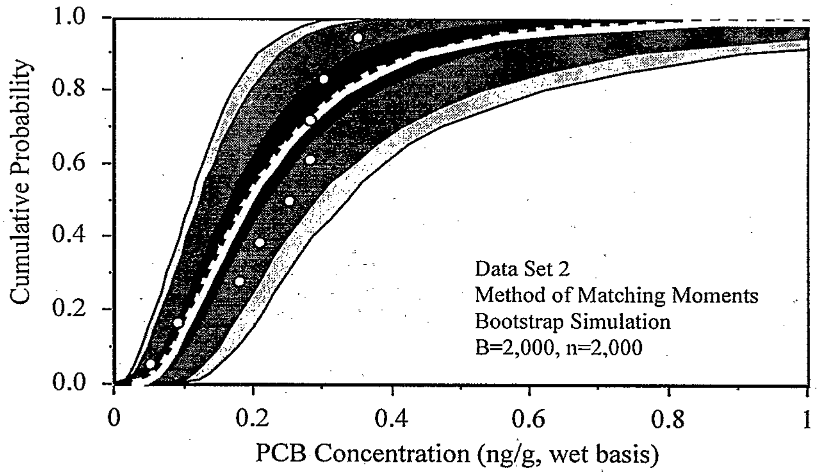
  While data may not be generated exactly according to a parametric distribution, evaluating parametric distributions may provide insight to generalizable features of a data set, such as moments, parameter values, or other statistics. Before deciding which distribution to use, two experts pointed out the value of trying to fit a parametric distribution to gain some insight about the data set (e.g., how particular parameters may be related to other aspects of the data set). These experts felt there is great value in examining larger data sets and thinking about what tools can be used to put data into better perspective. Another expert noted that the PDF is easier to defend at a public meeting or in a legal setting because it has some theoretical basis.

- *Assessing risk assessment outcome.* The importance of understanding what the implications of the distribution choice are to the outcome of the risk assessment was stressed. An example of fitting soil ingestion data to a number of parametric and non-parametric distributions yielded very different results. Depending on which distribution was used, cleanup goals were changed by approximately 2 to 3 times. Therefore, the choice may have cost implications.

- *Assuming all data are empirical.* One expert felt strongly that all distributions are empirical. In data poor situations, why assume that the data are Lognormal? The data could be bimodal in the tails. If a data set is assumed to be empirical, there is some control as to how to study the tails. Another expert reiterated that using EDFs in data poor situations (e.g., six data points) does not enable simulation above or below known data values. One expert disagreed providing an example that legitimizes the concern for assuming that data fit a parametric distribution. He noted that if there is no mechanistic basis for fitting a parametric distribution, and a small set of data points by chance are at the lower end of the distribution, the 90th percentile estimate will be wrong.

- *Evaluating uncertainty*. Techniques for estimating uncertainty in EDFs and PDFs were discussed. The workshop chair presented an example in which he fit a distribution for variability to nine data points. He then placed uncertainty bands around the distributions (both Normal and Lognormal curves) using parametric bootstrap simulation. (See Figure 6-1). For example, bands were produced by plotting the results of 2,000 runs of a synthetic data set of nine points sampled randomly from the Lognormal distribution fitted to the original data set. The wide uncertainty (probability) bands indicate the confidence in the distribution. This is one approach for quantifying how much is known about what is going on at the tails, based on random sampling error. When this exercise was performed for the Normal distribution, less uncertainty was predicted in the upper tail; however, a lower tail with negative values was predicted, which is not appropriate for a non-negative physical quantity such as concentrations. The chair noted that, if a stepwise EDF had been used, high and low ends would be truncated and tail concentrations would not have been predicted. This illustrates that the estimate of uncertainty in the tails depends on which assumption is made for the underlying distribution. Considering uncertainty in this manner allows the assessor to evaluate alternative distributions and gain insight on distinguishing between variability and uncertainty in a "2-dimensional probabilistic framework." Several participants noted that this was a valuable example.

Figure 6-1: Variability and Uncertainty in the Fit of Lognormal Distribution to a Data Set of n=9 (Frey, H.C. and D.E. Burmaster, 1998)



- *Extrapolating beyond the range of observable data.* The purpose of the risk analysis drives what assessors must know about the tails of the distribution. One expert emphasized that the further assessors go into the tails, the less they know. Another stressed that once assessors get outside the range of the data, they know nothing. Another expert disagreed with the point that assessors know nothing beyond the highest data point. He suggested using analogous data sets that are more data rich to help in predicting the tails of the distribution. The primary issue becomes how much the assessors are willing to extrapolate.

  Several experts agreed that uncertainty in the tails is not always problematic. If the assessor wants to focus on a subgroup, for example, it is not necessary to look at the tail of the larger group. *Stratification*, used routinely by epidemiologist, was suggested. With stratification, the assessor would look at the subgroup and avoid having to perform an exhaustive assessment of the tail, especially for more preliminary calculations used in a tiered approach. In a tiered risk assessment system, if the assessor assumes the data are Lognormal, standard multiplicative equations can be run on a simple calculator. While Monte Carlo-type analyses can provide valuable information in many cases, several experts agreed that probabilistic analyses are not always appropriate or necessary. It was suggested that, in some cases, deterministic scenario-based analyses, rather than Monte Carlo simulation, would be a useful way to evaluate extreme values for a model output.

  In a situation where a model is used to make predictions of some distribution, several experts agreed that the absence of perfect information about the tails of the distribution of each input does not mean that assessors will not have adequate information about the tail of the model output. Even if all we have is good information about the central portions of the input distributions, it may be possible to simulate an extreme value for the model output.

- *Use of data in the tails of the distribution.* One expert cautioned assessors to be sensitive to potentially important data in the tails. He provided an example in which assessors relied on the "expert judgement" of facility operators in predicting contaminant releases from a source. They failed to adequately predict "blips" that were later shown to exist in 20 to 30 percent of the distribution. Another expert noted that he was skeptical about adding tails (but was not skeptical about setting upper and lower bounds). It was agreed that, in general, assessors need to carefully consider what they do know about a given data set that could enable them to set a realistic upper bound (e.g., body weight). The goal is to provide the risk manager with an "unbiased estimate of risk." One expert reiterated that subjective judgments are inherent in the risk assessment process. In the case of truncating data, such judgments must be explained clearly and justified to the risk manager. In contrast to truncation, one expert reminded the group that the risk manager decides upon what percentile of the tail is of interest. Because situations arise in which the risk manager may be looking for 90th to 99th percentile values, the assessor must know how to approach the problem and, ultimately, must clearly communicate the approach and the possible large uncertainties.

- *Scenarios.* The group discussed approaches for evaluating the *high ends of distributions* (e.g., the treatment blips mentioned previously or the pica child). Should the strategy for assessing overall risks include high end or unusual behavior? Several experts felt that including extreme values in the overall distribution was not justified and suggested that the upper bounds in these cases be considered "scenarios." As with upper bounds, one expert noted that low end values also need special attention in some cases (e.g., air exchange in a tight house).

- *Generalized distributions versus mixtures.* Expert opinion differed regarding the issue of generalized versus mixture distributions. One expert was troubled by the notion of a mixture distribution. He would rather use a more sophisticated generalized distribution. Another expert provided an example of radon, stating that it is likely a mixture of Normal distributions, not a Lognormal distribution. Therefore, treatment of mixtures might be a reasonable approach. Otherwise, assessors risk grossly underestimating risk in concentrated areas by thinking they know the parametric form of the underlying distribution.

  The same expert noted that the issue of mixtures highlights the importance of having some theoretical basis for applying available techniques (e.g., possible Bayesian methods). Another expert stated that he could justify using distributions that are mixtures, because in reality many data sets are inherently mixtures.

- *Truncation of distributions*. Mixed opinions were voiced on this issue. One expert noted that assessors can *extend a distribution to a plausible upper bound* (e.g., assessors can predict air exchange rates because they know at a certain point they will not go higher). Another expert noted that truncating the distribution by 2 or 3 standard deviations is not uncommon because, for example, the assessors simply do not want to generate 1,500-pound people. One individual questioned, however, whether truncating a Lognormal distribution invalidates calling the distribution Lognormal. Another commented on instances in which truncating the distribution may be problematic. For example, some relevant data may be rejected. Also, the need to truncate suggests that the fit is very poor. The only reason to truncate, in his opinion, is if one is concerned about getting a zero or negative value, or perhaps an extremely high outlier value. One expert noted that truncation clearly has a role, especially when a strong scientific or engineering basis can be demonstrated.

- *When should neither an EDF nor PDF be used?* Neither an EDF nor a PDF may be useful/appropriate when large extrapolations are needed or when the assessor is uncomfortable with extrapolation beyond the available data points. In these cases, scenario analyses may come into play.

In their final discussions on EDF/PDF, the group widely encouraged *visual or graphical representation* of data. Additional thoughts on visually plotting the data are presented in the following discussions of goodness of fit.

6.2 GOODNESS-OF-FIT (GoF)

On what basis should it be decided whether a data set is adequately represented by a fitted parametric distribution?

The final workshop discussions related to the appropriateness of using available GoF test statistics in evaluating how well a data set is represented by a fitted distribution. Experts were asked to consider what options are best suited and how one chooses among multiple tests that may provide different answers. The following highlights the major points of these discussions.

- *Interpreting poor fit.* GoF in the middle of the distribution is not as important as that of the tails (upper and lower percentiles). Poor fit may be due to outliers at the other end of the distribution. If there are even only a few outliers, GoF tests may provide the wrong answer.

- *Graphical representation* of data is key to evaluating goodness or quality of fit. Unanimously, the experts agreed that using probability plots (e.g., EDF, QQ plots, PP plots) or other visual techniques in evaluating goodness of fit is an acceptable and recommended approach. In fact, the group felt that graphical methods should always be used. Generally, it is easier to judge the quality of fit using probability plots that compare data to a straight line. There may be cases in which a fit is rejected by a particular GoF test but appears reasonable when using visual techniques.

  The group supported the idea that GoF tests should not be the only consideration in fitting a distribution to data. Decisions can be made based on visual inspection of the data. It was noted that graphical presentations help to show quirks in the data (e.g., mixture distributions). It was also recommended that the assessor seek the consensus of a few trained individuals when interpreting data plots (as is done in the medical community when visually inspecting X-rays or CAT scans).

- *What is the significance of failing a weak test such as chi-square? Can we justify using data that fail a GoF test?* GoF tests may be sensitive to imperfections in the fit that are not important to the assessor or decision maker. The group therefore agreed that the fitted distribution can be used especially if the failure of the test is due to some part of the distribution that does not matter to the analysis (e.g., the lower end of the distribution). The reason the test failed, however, must be explained by the assessor. Failing a chi-square test is not problematic if the lower end of the distribution is the reason for the failure. One expert questioned whether the assessor could defend (in court) a failed statistical test. Another expert responded indicating that a graphical presentation might be used to defend use of the data, showing, for example, that the poor fit was a result of data set size, not chance.

- Considerations for risk assessors when GoF tests are used.

  — The evaluation of distributions is an estimation process (e.g., PDFs). Using a systematic testing approach based on the straight line null hypothesis may be problematic.

  — R^2 is a poor way to assess GoF.

  — The appropriate loss function must be identified.

— The significance level must be determined before the data are analyzed. Otherwise, it is meaningless. It is a risk management decision. The risk assessor and risk manager must speak early in the process. The risk manager must understand the significance level and its application.

- *Should GoF tests be used for parameter estimation* (e.g., objective function is to minimize the one-tail Anderson-Darling)? A degree of freedom correction is needed before the analysis is run. The basis for the fit must be clearly defined—are the objective and loss functions appropriate?

- "Maximum likelihood estimation (MLE)" is a well-established statistical tool and provides a relatively easy path for separating variability from uncertainty.

- The adequacy of Crystal Ball®'s curve-fitting capabilities was questioned. One of the experts explained that it runs three tests, then ranks them. If the assessor takes this one step further by calculating percentiles and setting up plots, it is an adequate tool.

- The Office of Water collects large data sets. Some of the office's efforts might provide some useful lessons into interpreting data in the context of this workshop.

- *What do we do if only summary statistics are available?* Summary statistics are often all that are available for certain data sets. The group agreed that MLE can be used to estimate distribution parameters from summary data. In addition, one expert noted that probability plots are somewhat useful for evaluating percentile data. Probability plots enable assessors to evaluate the slope (standard deviation) and the intercept (mean). Confidence intervals cannot be examined and uncertainty cannot be separated from variability.

In summary, the group identified possible weaknesses associated with using statistical GoF tests in the context described above. The experts agreed unanimously that graphical/visual techniques to evaluate how well data fit a given distribution (alone or in combination with GoF techniques) may be more useful than using GoF techniques alone.

6.3 SUMMARY OF EDF/PDF AND GoF DISCUSSIONS

The experts agreed, in general, that the choice of an EDF versus a PDF is a matter of personal preference. The group recommended, therefore, that no rigid guidance be developed requiring one or the other in a particular situation. The decision on which distribution function to use is dependent on several factors, including the number of data points, the outcome of interest, and how interested the assessor is in the tails of the distribution. Varied opinions were voiced on the use of mixture distributions and the appropriateness of truncating distributions. The use of scenario analysis was suggested as an alternative to probabilistic analysis when a particular input cannot be assigned a probability distribution or when estimating the tails of an important input distribution may be difficult.

Regarding GoF, the group fully agreed that visualization/graphic representation of both the data and the fitted distribution is the most appropriate and useful approach for ascertaining adequacy of fit. In general, the group agreed that conventional GoF tests have significant shortcomings and should not be the primary method for determining adequacy of fit.

SECTION SEVEN

OBSERVER COMMENTS

This section presents observers' comments and questions during the workshop, as well as responses from the experts participating in the workshop.

DAY ONE: Tuesday, April 21, 1998

Comment 1
Helen Chernoff, TAMS Consultants

Helen Chernoff said that, with the release of the new policy, users are interested in guidance on how to apply the information on data representativeness and other issues related to probabilistic risk assessment. She had believed that the workshop would focus more on application, rather than just on the background issues of probabilistic assessments. What methods could be used to adjust data and improve data representativeness (e.g., the difference between past and current data usage)?

Response

The workshop chair noted that adjustment discussions during the second day of the workshop start to explore available methods. One expert stated that, based on his impression, the workshop was designed to gather input from experts in the field of risk assessment and probabilistic techniques. He noted that EPA's policy on probabilistic analysis emerged only after the 1996 workshop on Monte Carlo analysis. Similarly, EPA will use the information from this workshop to help build future guidance on probabilistic techniques, but EPA will not release specific guidance immediately (there may be an approximate two-year lag).

The chair noted that assessors may want to know when they can/should implement alternate approaches. He pointed out that the representativeness issue is not specific to probabilistic assessment. It applies to all assessments. Since EPA released its May 1997 policy on Monte Carlo analysis, representativeness has been emphasized more, especially in exposure factor and distribution evaluations. He noted, however, that data quality/representativeness is equally important when considering a point estimate. However, it may not be as important if a point estimate is based on central tendency instead of an upper percentile where there may be fewer data. Another agreed that the representativeness issue is more important for probabilistic risk assessment than deterministic risk assessment (especially a point estimate based on central tendency).

Comment 2
Emran Dawoud, Human Health Risk Assessor, Oak Ridge National Laboratory

Mr. Dawoud commented that the representativeness question should reflect whether additional data must be collected. He noted that the investment (cost/benefit) should be considered. From a risk assessment point of view, one must know how more data will affect the type or cost of remedial activity. In his opinion, if representativeness does not change the type or cost of remedial activity, further data collection is unwarranted.

Mr. Dawoud also commented that the risk model has three components: source, exposure, and dose-response. Has the sensitivity of exposure component been measured relative to the sensitivity of the other two components? He noted the importance of the sensitivity of the source term, especially if fate and transport are involved.

Mr. Dawoud briefly noted that, in practice, a Lognormal distribution is being fit with only a few samples. Uncertainty of the source term in these cases is not quantified or incorporated into risk predictions. Even if standard deviation is noted, the contribution to final risk prediction is not considered. Mr. Dawoud noted that the workshop discussions on the distribution around exposure parameters seem to be less important than variation around the source term. Likewise, he noted the uncertainties associated with the dose-response assessment as well (e.g., applying uncertainty factors of 10, 100, etc.).

Response

One participant noted that representativeness involves more than collecting more data. Evaluating representativeness is often about choosing from several data sets. He agreed that additional data are collected depending on how the collection efforts may affect the bottom line assessment answer. He noted that if input does not affect output, then its distribution need not be described.

Relative to Mr. Dawoud's second point, it was noted that source term evaluation is part of exposure assessment. While exposure factors (e.g, soil ingestion and exposure duration) affect the risk assessment, one expert emphasized that the most important driving "factor" is the source term. As for dose-response, the industry is just beginning to explore how to quantify variability and uncertainty.

The workshop chair noted that methodologically, exposure and source terms are not markedly different. The source term has representativeness issues. There are ways to distinguish between variability and uncertainty in the variability estimate.

Lastly, more than one expert agreed that the prediction of risk for noncancer and cancer endpoints (based on the reference dose [RfD] and cancer slope factor [CSF], respectively) is very uncertain. The methods discussed during this workshop cannot be directly applied to RfDs and CSFs, but they could be used on other toxicologic data. More research is needed in this area.

Comment 3
Ed Garvey, TAMS Consultants

Mr. Garvey questioned whether examining factors of 2 or 3 on the exposure side is worthwhile, given the level of uncertainty on the source or dose-response term, which can be orders of magnitude.

Response

It was an EPA policy choice to examine distributions looking first at exposure parameters, according to one EPA panelist. He also reiterated that the evaluation of exposure includes the source term (i.e., exposure = concentration x uptake/averaging time). One person noted that it was time to "step up" on quantifying toxicity uncertainty. Exposure issues have been driven primarily by engineering approaches (e.g., the Gaussian plume model), toxicity has historically been driven by toxicologists and statisticians and are more data oriented.

It was noted that, realistically, probabilistic risk assessments will be seen only when money is available to support the extra effort. Otherwise, 95% UCL concentrations and point estimates will continue to be used. Knowing that probabilistic techniques will enable better evaluations of variability and uncertainty, risk assessors must be explicitly encouraged to perform probabilistic assessments. We must accept that the existing approach to toxicity assessment, while lacking somewhat in scientific integrity, is the only option at present.

Comment 4
Emran Dawoud, Human Health Risk Assessor, Oak Ridge National Laboratory

Mr. Dawoud asked whether uncertainty analysis should be performed to evaluate fate and transport related estimates.

Response

One expert stressed that whenever direct measurements are not available, variability must be assessed. He commented that EPA's Probabilistic Risk Assessment Work Group is preparing two chapters for Risk Assessment Guidance for Superfund (RAGS): one on source term variability and another on time-dependent considerations of the source term.

Comment 5
Zubair Saleem, Office of Solid Waste, U.S. EPA

Mr. Saleem stated that he would like to reinforce certain workshop discussions. He commented that any guidance on probabilistic assessments should not be too rigid. Guidance should clearly state that methodology is evolving and may be revised. Also, guidance users should be encouraged to collect additional data.

Response

The workshop chair recognized Mr. Saleem's comment, but noted that the experts participating in the workshop can only provide input and advice on methods, and is not in a position to recommend specific guidelines to EPA.

DAY TWO: Wednesday, April 22, 1998

Comment 1
Lawrence Myers, Research Triangle Institute

Mr. Myers offered a word of caution regarding GoF tests. He agrees that many options do not work well but he stated that in an adversarial situation (e.g., a court room) he would rather be defending data distributions based on a quantitative model instead of a graphical representation.

Mr. Myers noted that the problem with goodness of fit is the tightness of the null hypothesis (i.e., it specifies that the true model is exactly a member of the particular class being examined). Mr. Myers cited Hodges and Layman (1950s) who generalized chi-square in a way that may be meaningful to the issues discussed in this workshop. Specifically, because exact conformity is not expected, a more appropriate null hypothesis would be that the true distribution is "sufficiently close" to the family being examined.

Response

One expert reiterated that when a PDF is fitted, it is recognizably an approximation and therefore makes application of standard GoF statistics difficult. Another expressed concern that practitioners could go on a "fishing expedition," especially in an adversarial situation, to find a GoF test that gives the right answer. He did not feel this is the message we want to be giving practitioners. A third expert noted a definite trend in the scientific community away from GoF tests and towards visualization.

SECTION EIGHT

REFERENCES

Buck, R.J., K.A. Hammerstrom, and P.B. Ryan, 1995. Estimating Long-Term Exposures from Short-term Measurements. *Journal of Exposure Analysis and Environmental Epidemiology*, Vol. 5, No. 3, pp. 359-373.

Efron, B. and R.J. Tibshirani, 1993. An Introduction to the Bootstrap. Chapman and Hall. New York.

Frey, H.C. and D.E. Burmaster, "Methods for Characterizing Variability and Uncertainty: Comparison of Bootstrap Simulation and Likelihood-Based Approaches," *Risk Analysis* (Accepted 1998).

RTI, 1998. *Development of Statistical Distributions for Exposure Factors*. Final Report. Prepared by Research Triangle Institute. U.S. EPA Contract 68D40091, Work Assignment 97-12. March 18, 1998.

U.S. Environmental Protection Agency, 1996a. Office of Research and Development, National Center for Environmental Assessment. *Exposure Factors Handbook*, SAB Review Draft (EPA/600/P-95/002Ba).

U.S. Environmental Protection Agency, 1996b. *Summary Report for the Workshop on Monte Carlo Analysis*. EPA/630/R096/010. September 1996.

U.S. Environmental Protection Agency, 1997a. *Guiding Principles for Monte Carlo Analysis*. EPA/630/R-97/001. March 1997.

U.S. Environmental Protection Agency, 1997b. Policy for Use of Probabilistic Analysis in Risk Assessment at the U.S. Environmental Protection Agency. May 15, 1997.

Wallace, L.A., N. Duan, and R. Ziegenfus, 1994. Can Long-term Exposure Distributions Be Predicted from Short-term Measurements? *Risk Analysis*, Vol. 14, No. 1, pp. 75-85.

APPENDIX A

ISSUE PAPERS

Issue Paper on Evaluating Representativeness of Exposure Factors Data

This paper is based on the Technical Memorandum dated March 4, 1998, submitted by Research Triangle Institute under U.S. EPA contract 68D40091.

1. INTRODUCTION

The purpose of this document is to discuss the concept of representativeness as it relates to assessing human exposures to environmental contaminants and to factors that affect exposures and that may be used in a risk assessment. (The factors, referred to as exposure factors, consist of measures like tapwater intake rates, or the amount of time that people spend in a given microenvironment.) This is an extremely broad topic, but the intent of this document is to provide a useful starting point for discussing this extremely important concept.

Section 2 furnishes some general definitions and notions of representativeness. Section 3 indicates a general framework for making inferences. Components of representativeness are presented in Section 4, along with some checklists of questions that can help in the evaluation of representativeness in the context of exposures and exposure factors. Section 5 presents some techniques that may be used to improve representativeness. Section 6 provides our summary and conclusions.

2. GENERAL DEFINITIONS/NOTIONS OF REPRESENTATIVENESS

Representativeness is defined in *American National Standard: Specifications and Guidelines for Quality Systems for Environmental Data and Environmental Technology Programs (ANSI/ASQC E4 - 1994)* as follows:

> The measure of the degree to which data accurately and precisely represent a characteristic of a population, parameter variations at a sampling point, a process condition, or an environmental condition.

Although Kendall and Buckland (*A Dictionary of Statistical Terms*, 1971) do not define representativeness, they do indicate that the term "representative sample" involves some confusion about whether this term refers to a sample "selected by some process which gives all samples an equal chance of appearing to represent the population" or to a sample that is "typical in respect of certain characteristics, however chosen." Kruskal and Mosteller (1979) point out that representativeness does not have an unambiguous definition; in a series of three papers, they present and discuss various notions of representativeness in the scientific, statistical, and other literature, with the intent of clarifying the technical meaning of the term.

In Chapter 1 of the *Exposure Factors Handbook* (EFH), the considerations for including the particular source studies are enumerated and then these considerations are evaluated qualitatively at the end of each chapter (i.e., for each type of exposure factor data). One of the criteria is "representativeness of the population," although there are several other criteria that clearly relate to various aspects of representativeness. For example, these related criteria include the following:

EFH Study Selection Criterion	EFH Perspective
focus on factor of interest	studies with this specific focus are preferred
data pertinent to U.S.	studies of U.S. residents are preferred
current information	recent studies are preferred, especially if changes over time are expected
adequacy of data collection period	generally the goal is to characterize long-term behavior
validity of approach	direct measurements are preferred
representativeness of the population	U.S. national studies are preferred
variability in the population	studies with adequate characterizations of variability are desirable
minimal (or defined) bias in study design	studies having designs with minimal bias are preferred (or with known direction of bias)
minimal (or defined) uncertainty in the data	large studies with high ratings on the above considerations are preferred

3. A GENERAL FRAMEWORK FOR MAKING INFERENCES

Despite the lack of specificity of a definition of representativeness, it is clear in the present context that representativeness relates to the "comfort" with which one can draw inferences from some set(s) of extant data to the population of interest for which the assessment is to be conducted, and in particular, to certain characteristics of that population's exposure or exposure factor distribution. The following subsections provide some definitions of terms and attempt to break down the overall inference into some meaningful steps.

3.1 Inferences from a Sample to a Population

In this paper, the word **population** to refers to a set of units which may be defined in terms of person and/or space and/or time characteristics. The population can thus be defined in terms of its individuals' characteristics (defined by demographic and socioeconomic factors, human behavior, and study design) (e.g., all persons aged 16 and over), the spatial characteristics (e.g., living in Chicago) and/or the temporal characteristics (e.g., during 1997).

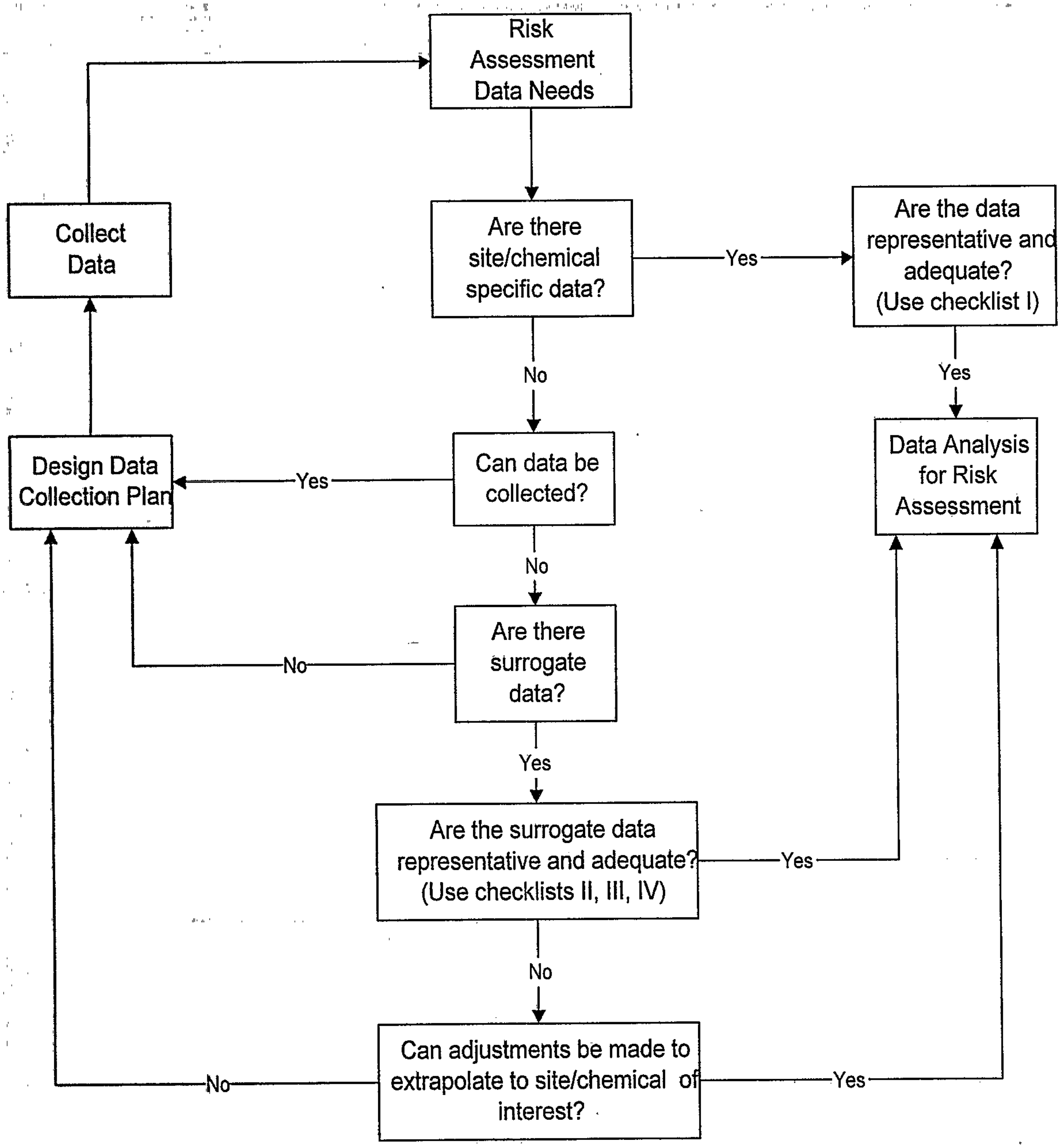
In conducting a risk assessment, the assessor needs to define the population of concern — that is, the set of units for which risks are to be assessed (e.g., lifetime risks of all U.S. residents). At a Superfund site, this population of concern is generally the population surrounding the site. In this document, the term population of concern refers to that population for which the assessor wishes to draw inferences. If it were practical, this is the population for which a census (a 100% sample) would exist or for which the assessor would conduct a probability-based study of exposures. Figure 1 provides a diagram of the exposure assessor decision process during the selection of data for an exposure assessment.

As depicted in figure 1, quite often it is not practical or feasible to obtain data on the population of concern and the assessor has to rely on the use of surrogate data. These data generally come from studies conducted by researchers for a variety of purposes. Therefore, the assessor's population of concern may differ from the surrogate population. Note that the population differences may be in any one (or more) of the characteristics described earlier. For example, the surrogate population may only cover a subset of the individuals in the assessor's population of concern (Chicago residents rather than U.S. residents). Similarly, the surrogate data may have been collected during a short period of time (e.g., days), while the assessor may be concern about chronic exposures (i.e., temporal characteristics).

The studies used to derive these surrogate data are generally designed with a population in mind. Since it may not be practical to sample everyone in that population, probability-based sampling are often conducted. This sampling scheme allows valid statistical (i.e., non-model-based) inferences, assuming there were no implementation difficulties (e.g., no nonresponse and valid measurements). Ideally, the implementation difficulties would not be severe (and hence ignored), so that these sampled individuals can be considered representative of the population. If there are implementation difficulties, adjustments are typically made (e.g., for nonresponse) to compensate for the population differences. Such remedies for overcoming inferential gaps are fairly well documented in the literature in the context of probability-based survey sampling (e.g., see Oh and Scheuren (1983)). If probability sampling is not employed, the relationships of the selected individuals for which data are sought and of the respondents for which data are actually acquired to the population for which the study was designed to address are unclear.

There are cases where probability-based sampling is used and the study design allows some model-based inferences. For instance, food consumption data are often obtained using surveys which ask respondents to recall food eaten over a period of few days. These data are usually collected throughout a one-year period to account for some seasonal variation in food consumption. Statistical inferences can then be made for the individuals surveyed within the time frame of study. For example, one can estimate the mean, the 90th percentile, etc. for the number

Figure 1: Risk Assessment Data Collection Process


Risk Assessment Data Needs
Are there site/chemical specific data?
Yes
Are the data representative and adequate? (Use checklist I)
Yes
Data Analysis for Risk Assessment
No
Can data be collected?
Yes
Design Data Collection Plan
Collect Data
No
Are there surrogate data?
No
Yes
Are the surrogate data representative and adequate? (Use checklists II, III, IV)
Yes
No
Can adjustments be made to extrapolate to site/chemical of interest?
No
Yes

of days during which individuals were surveyed. However, if at least some of the selected individuals are surveyed multiple periods of time during that year, then a model-based strategy might allow estimation of a distribution of long-term (e.g., annual) consumption patterns.

If probability-based sampling is not used, model-based rather than statistical inferences are needed to extend the sample results to the population for which the study was designed.

In contrast to the inferences described above, which emanate from population differences and the sampling designs used in the study, there are two additional inferential aspects that relate to representativeness:

- The degree to which the study design is followed during its implementation
- The degree to which a measured value represents the true value for the measured unit

Both of these are components of measurement error. The first relates to an implementation error in which the unit selected for measurement is not precisely the one for which the measurement actually is made. For instance, the study's sampling design may call for people to record data for 24-hr periods starting at a given time of day, but there may be some departure from this ideal in the actual implementation. The second has to do with the inaccuracy in the measurement itself, such as recall difficulties for activities or imprecision in a personal air monitoring device.

4. COMPONENTS OF REPRESENTATIVENESS

As described above, the evaluation of how representative a data set is begins with a clear definition of the population of concern (the population of interest for the given assessment), with attention to all three fundamental characteristics of the population — individual, spacial, and temporal characteristics. Potential inferential gaps between the data set and the population of concern -- that is, potential sources of unrepresentativeness -- can then be partitioned both along these population characteristics. Components of representativeness are illustrated in Table1: the rows correspond to the inferential steps and the columns correspond to the population characteristics. The inferential steps are distinguished as being either internal or external to the source study.

4.1 Internal Components - Surrogate Data Versus the Study Population

After determining that a study provides information on the exposures or exposure factors of interest, it is important that the exposure assessor evaluate the representativeness of the surrogate study (or studies). This entails gaining an understanding of both the individuals sampled for the study and the degree to which the study achieved valid inferences to that population. The assessor should consider the questions in Checklist I in the appendix to help establish the degree of representativeness inherent to this internal component. In the context of the Exposure Factors Handbook (EFH), the representativeness issues listed in this checklist are presumably the types of considerations that led to selection of the source studies that appear in

Table 1. Elements of Representativeness

Component of Inference	Population Characteristics		
	Individual Characteristics	Spacial Characteristics	Temporal Characteristics
EXTERNAL TO STUDY			
How well does the surrogate population represent the population of concern?	• Exclusion or limited coverage of certain segments of population of concern	• Exclusion or inadequate coverage of certain regions or types of areas (e.g., rural areas) that make up the population of concern	• Lack of currency • Limited temporal coverage, including exclusion or inadequate coverage of seasons • Inappropriate duration for observations (e.g., short-term measurements where concern is on chronic exposures)
INTERNAL TO STUDY			
How well do the individuals sampled represent the population of concern for the study?	• Imposed constraints that exclude certain segments of study population • Frame inadequacy (e.g., due to lack of current frame information)	• Inadequate coverage (e.g., limited to single urban area)	• Limited temporal coverage (e.g., limited study duration) • Inappropriate duration for observations
How well do the actual number of respondents represent the sampled population? How well does the measured value represent the true value for the measured unit?	• Non-probability sample of persons • Excessive nonresponse • Inadequate sample size • Behavior changes resulting from participation in study (Hawthorne effect) • Measurement errors associated with people's ability/desire to respond accurately to questionnaire items • Measurement error associated with within-specimen heterogeneity • Inability to acquire physical specimen with exact size or shape or volume desired	• Non-probability sample of spatial units (e.g., convenience or judgmental siting of ambient monitors) • Inaccurate identification of sampled location	• Non-probability sample of observation times • Deviation in times selected vs. those measured or reported (e.g., due to schedule slippage, or incomplete response) • Measurement errors related to time (e.g., recall difficulties for foods consumed or times in microenvironments)

the EFH. As indicated previously, the focus for addressing representativeness in that context was national and long-term, which may or may not be consistent with the assessment of current interest.

4.2 External Components - Population of Concern Versus Surrogate Population

In many cases, the assessor will be faced with a situation in which the population of concern and surrogate population do not coincide in one or more aspects. To address this external factor of representativeness, the assessor needs to:

- determine the relationship between the two populations
- judge the importance of any discrepancies between the two populations
- assess whether adjustments can be made to reconcile or reduce differences.

To address these, the assessor needs to consider all characteristics of the populations. Relevant questions to consider are listed in Checklists II, III, and IV in the appendix for the individual, spacial, and temporal characteristics, respectively.

Each checklist contains several questions related to each of the above bullets. For example, the first few items of each checklist relate to the first item above (relationship of the two populations). There are several possible ways in which the two populations may relate to each other; these cases are listed below and can be addressed for each population dimension:

- Case 1: The population of concern and surrogate population are (essentially) the same
- Case 2: The population of concern is a subset of the surrogate population
  - Case 2a: The subset is a large and identifiable subset.
  - Case 2b: The subset is a small and/or unidentifiable subset.
- Case 3: The surrogate population is a subset of the population of concern.
- Case 4: The population of concern and surrogate population are disjoint.

Note that Case 2a implies that adequate data are available from the surrogate study to generate separate summary statistics (e.g., means, percentiles) for the population of concern. For example, if the population of concern was focused on children and the surrogate population was a census or large probability study of all U.S. residents, then children-specific summaries would be possible. In such a situation, Case 2a reverts to back to Case 1.

Case 2b will be typical of situations in which large-scale (e.g., national or regional) data are available but assessments are needed for local areas (or for acute exposures). As an example, suppose raw data from the National Food Consumption Survey (NFCS) can be used to form meaningful demographic subgroups and to estimate average tapwater consumption for such subgroups (e.g., see Section 5.1). If a risk assessment involving exposure from copper smelters is to be conducted for the southwestern U.S., for instance, tapwater consumption would probably be considered to be different for that area than for the U.S. as a whole, but the NFCS data for that area might be adequate. If so, this would be considered Case 2a. But if the risk assessment concerned workers at copper smelters, then an even greater discrepancy between the population of concern and the surrogate data might be expected, and the NFCS data would likely be regarded as inadequate, and more speculative estimates would be needed.

In contrast to Case 2, Case 3 will be typical of assessments that must use local and/or short-term data to extrapolate to regional or national scales and/or to long-term (chronic) exposures. Table 2 presents some hypothetical examples for each case. Note that, as illustrated here and as implied by the bulleted items in Checklist IV, the temporal characteristics has two series of issues: one that relates to the currency and the temporal coverage (study duration) of the source study relative to the population of concern time frame, and one that relates to the time unit of observation associated with the study.

Since most published references to the NFCS rely on the 1977-78 survey, exposure factor data based on that survey might well be considered as Case 4 with respect to temporal coverage, as trends such as consumption of bottled water and organic foods may not be well represented by 20 year-old data. A possible approach in this situation would be to obtain data from several NFCSs, to compare or test for a difference between them, and to use them to extrapolate to the present or future. The NFCS also illustrates the other temporal aspect — dealing with a time-unit mismatch of the data and the population of concern — since the survey involves three consecutive days for each person, while typically a longer-term estimate would be desired, e.g., a person-year estimate (e.g., see Section 5.2).

While determining the relationship of the two populations will generally be straightforward (first bullet), determining the importance of discrepancies and making adjustments (the second and third bullets) may be highly subjective and require an understanding of what factors contribute to heterogeneity in exposure factor values and speculation as to their influence on the exposure factor distribution. Cases 1 and 2a are the easiest, of course. In the other cases, it will generally be easier to speculate about how the mean and variability (perhaps expressed as a coefficient of variation (CV)) of the two populations may differ than to speculate on changes in a given percentile. Considerations of unaffected portions of the population must also be factored into the risk assessor's speculation. The difficulty in such speculation obviously increases dramatically when two or more factors affect heterogeneity, especially if the factors are anticipated to have opposite or dependent effects on the exposure factor values. Regardless of how such speculation is ultimately reflected in the assessment (either through ignoring the population differences or by adjusting estimated parameters of the study population), recognition of the increased uncertainty should be incorporated into sensitivity analyses. As a part of such an analysis, it would be instructive to determine risks, when, for each relevant factor (e.g., age category), several assessors independently speculate on the mean (e.g., a low, best guess, and high) and on the CV.

Table 2. Examples of Relationships Between the Population of Concern and the Surrogate Population

Population Characteristics	Population	Case 1: Population of concern and surrogate population are the same	Case 2a: Population of concern is a subset of the surrogate population, and data on subset are available	Case 2b: Population of concern is a subset of surrogate population and data on subset are not available	Case 3: Surrogate population is a subset of the population of concern	Case 4: Population of concern and surrogate population are disjoint
Individual Characteristics	Population of concern:	U.S. residents	U.S. children	Asthmatic U.S. children	U.S. residents	U.S. children
	Surrogate population:	U.S. residents	U.S. residents + age data	U.S. residents	U.S. adults	U.S. adults
Spacial Characteristics	Population of concern:	U.S.	Northeast U.S.	Near hazardous waste sites	U.S.	U.S.
	Surrogate population:	U.S.	U.S. + region ID data	U.S.	Chicago	Netherlands
Temporal Characteristics and Currency	Population of concern:	one year, 1998	summer, 1998	1998 days with smog	lifetime	future years
	Surrogate population:	one year, 1998	one year, 1998 + season ID data	one year, 1998	two summertime weeks, 1998	1996
Temporal Observation Units	Population of concern:	person-days	eating occasions	eating occasions (acute)	lifetimes (chronic)	NA
	Surrogate population:	person-days	person-days + meal-specific data	person-3-day data	person-days	

5. ATTEMPTING TO IMPROVE REPRESENTATIVENESS

5.1 Adjustments to Account for Differences in Population Characteristics or Coverage.

If there is some overlap in information available for the population of concern and the surrogate population (e.g., age distributions), then adjustments to the sample data can be made that attempt to reduce the bias that would result from directly applying the study results to the population of concern. Such methods of adjustment can all be generally characterized as "direct standardization" techniques, but the specific methodology to use depends on whether one has access to the raw data or only to summary statistics, as is often the case when using data from the Exposure Factors Handbook. With access to the raw data, the applicable techniques also depend on whether one wants to standardize to a single known population of concern distribution (e.g., age categories), to two or more marginal distributions known for the population of concern, or even to population of concern totals for continuous variables.

Summary Statistics Available. Suppose that the available data are summary statistics such as the mean, standard deviation, and various percentiles for an exposure factor of interest (e.g., daily consumption of tap water). Furthermore, suppose that these statistics are available for subgroups based on age, say age groups g = 1, 2, ..., G. Furthermore, suppose we know that the age distribution of the population of concern differs from that represented by the sample data. We can then estimate linear characteristics of the population of concern, such as the mean or the proportion exceeding a fixed threshold, using a simple weighted average. For example, the mean of the population of concern can be estimated as

$$\bar{x}_{ATP} = \Sigma_g \, P_g \bar{x}_g,$$

where Σ_g represents summation over the population of concern groups indexed by g, P_g is the proportion of the population of concern that belongs to group g, and $\bar{x}_g$ is the sample mean for group g.

Unfortunately, if one is interested in estimating a non-linear statistic for the population of concern, such as the variance or a percentile, this technique is not algebraically correct. However, lacking any other information from the sample, calculating this type of weighted average to estimate a non-linear population of concern characteristic is better than making no adjustment at all for known population differences. In the case of the population variance, we recommend calculating the weighted average of the group standard deviations, rather than their variances, and then squaring the estimated population of concern standard deviation to get the estimated population of concern variance.

Raw Data Available. If one has access to the raw data, not just summary statistics, options for standardization are more numerous and can be made more rigorously. The options depend, in part, on whether or not the data already have statistical analysis weights, such as those appropriate for analysis of data from a probability-based sample survey.

Suppose that one has access to the raw data from a census or from a sample in which all units can be regarded as having been selected with equal probabilities (e.g., a simple random sample). In this case, if one knows the number, N_g, of population of concern members in group g, then the statistical analysis weight to associate with the i-th member of the g-th group is

$$W_g(i) = \frac{N_g}{n_g},$$

where the sample contains n_g members of group g. Alternatively, if one knows only the proportion of the population and sample that belong to each group, one can calculate the weights as

$$W_g(i) = \frac{P_g}{p_g},$$

where p_g is the proportion of the sample in group g. The latter weights differ from those above only by a constant, the reciprocal of the sampling fraction, and will produce equivalent results for means and proportions. However, the former weights must be used to estimate population totals. In either case, the population of concern mean can be estimated as

$$\bar{x}_{ATP} = \frac{\Sigma_g \Sigma_i \; W_g(i) \, x_g(i)}{\Sigma_g \Sigma_i \; W_g(i)},$$

where $x_g(i)$ is the value of the characteristic of interest (e.g., daily tap water consumption) for the i-th sample member in group g.

In general, one may have access to weighted survey data, such as results from a probability-based sample of the surrogate population. In this case, the survey analysis weight, $w(i)$, for the i-th sample member is the reciprocal of that person's probability of selection with appropriate adjustments to reduce nonreponse bias and other potential sources of bias with respect to the surrogate population. Further adjustments for making inferences to the population of concern are considered below. These results can also be applied to the case of equally weighted survey data, considered above, by considering the survey analysis weight, $w(i)$, to be unity (1.00) for each sample member.

If one knows the distribution of the population of concern with respect to a given characteristic (e.g., the age/race/gender distribution), then one can use the statistical technique of poststratification to adjust the survey data to provide estimates adjusted to that same population distribution (see, e.g., Holt and Smith, 1979).[1] In this case, the weight adjustment factor for each member of poststratum g is calculated as

[1] Sampling variances are computed differently for standardized and poststratified estimates, but these details are suppressed in the present discussion (see, e.g., Shah *et al.*, 1993).

$$A_g = \frac{N_g}{\Sigma_{i\in g}\, w(i)},$$

where the summation is over all sample members belonging to poststratum *g*. The poststratified analysis weight for the *i*-th sample member belonging to poststratum *g* is then calculated as

$$w_p(i) = A_g\, w(i).$$

Using this weight, instead of the surrogate population weight, *w(i)*, standardizes the survey estimates to the population of concern.

If one knows multiple marginal distributions for the population of concern but not their joint distribution (e.g., marginal age, race, and gender distributions), one can apply a statistical weight adjustment procedure known as raking, or iterative proportional fitting, to standardize the survey weights (see, e.g., Oh and Scheuren, 1983). Raking is an iterative procedure for scaling the survey weights to known marginal totals.

If one knows population of concern subgroup totals for continuous variables, a generalized raking procedure can be used to standardize the survey weights to known distributions of categorical variables as well as known totals for continuous variables. The generalized raking procedures utilize non-linear, exponential modeling (see, e.g., Folsom, 1991 and Deville *et al.*, 1993).

Of course, none of these standardization procedures results in inferences to the population of concern that are as defensible as those from a well-designed sample survey selected from a sampling frame that completely and adequately covers the population of concern.

5.2 Adjustments to Account for Time-Unit Differences.

A common way in which the surrogate population and population of concern may differ is in the time unit of (desired) observation. Probably the most common situation occurs when the study data represent short-term measurements but where chronic exposures are of interest. In this case, some type of model is needed to make the time-unit inference (e.g., from the distribution of person-day or person-week exposures to the distribution of annual or lifetime exposures). In general, it is convenient to break down the overall inference into two components: from the time unit of measurement to the time duration of the study (data to the surrogate population), and from the time duration of the surrogate population to the time unit of the population of concern. For specificity, let t denote the observation time (e.g., a day or a week); let τ denote the duration of the study (i.e., τ is the time duration associated with the surrogate population); and let T denote the time unit of the population of concern (e.g., a lifetime). In the case of chronic exposure concerns, $t<\tau<T$.

Suppose that N denotes the number of persons in the surrogate population, and assume there are (conceptually) K disjoint time intervals of length t that surrogate population τ (i.e., Kt=τ). Thus a census of the surrogate population would involve NK short-term measurements (of exposures or of exposure factors). This can be viewed as a two-way array with N rows (persons) and K columns (time periods). Clearly, the distribution of these NK measurements, whose mean is the grand total over the NK cells divided by NK, encompasses both variability among people and variability among time periods within people (and in practice, measurement error also). The average across the columns for a given row (the marginal mean) is the average exposure for the given person over a period of length τ. Since the mean of these τ-period "measurements" over the N rows leads to the same mean as before, it is clear that the mean of the t-time measurements and the mean of the τ-time measurements is the same. However, unless there is no within-person variability, the variability of the longer τ-period measurements will be smaller than the variability of the shorter t-period measurements. If the distribution of the shorter term measurements is right-skewed, as is common, then one would expect the longer term distribution to exhibit less skewness. Note that the degree to which the variability shrinks depends on the relation between the within-person and between-person components of variance, which is related to the temporal correlation. For example, if there is little within-person variability, then people with high (low) values will remain high (low) over time, implying that the autocorrelation is high and that the shrinkage in variability in going from days to years (say) will be minimal. If there is substantial within-person variation, then the autocorrelations will be low and substantial shrinkage in the within-person variance (on the order of a t/τ decrease) will occur.

To make this t-to-τ portion of the inference, we therefore would ideally have a valid probability-based sample of the NK person-periods, and data on the t-period exposures or exposure factors would be available for each of these sampling units. As a part of this study design, we would also want to ensure that at least some of persons have measurements for more than one time period, since models that allow the time extrapolation will need data that, in essence, will support the estimation of within-person components of variability. There are several examples of models of this sort, some of which are described below.

Wallace *et al.* (1994) describe a model, which we refer to as the Duan-Wallace (DW) model, in which data over periods of length t, 2t, 3t, etc. (i.e., over any averaging period of length mt) are all conceptually regarded to be approximated by lognormal distributions, with parameters that depend on a "lifetime" variance component and a short term variance component. While such an assumption is theoretically inconsistent if exact lognormality is required, it may nevertheless serve well as an approximation. The basic notion of the DW method is that, while the mean of the exposures stays constant, the variability decreases as the number of periods averaged together increases. Hence it is assumed that the total variability for a distribution that averages over M periods (M=1,2,...) can be expressed in terms of a long-term component and a short-term component. Let γ_L and γ_S denote, respectively, the log-scale variances for these two components. Under the lognormal model, Wallace *et al.* show that the log-scale variance for the M-period distribution (i.e., the distribution that averages over M periods) is given by

$$V_M = \gamma_L + \log[1 + \frac{\exp(\gamma_S) - 1}{M}].$$

Note that an implication of the DW model is that the geometric means for the various distributions will increase as M increases. In fact, the geometric mean (gm) associated with the average of M short-term measurements will be

$$gm(M) = \bar{Y} \exp[-V_M /2]$$

where $\bar{Y}$is the overall population mean of the exposures. As a consequence, if data are adequate for estimating the variance components (and the mean of the exposures), then an estimated distribution for any averaging time can be inferred. In particular, the DW method can be applied if data are available for estimating V_M for (at least) two values of M, since one is then able to determine values of the two variance components. For instance, if two observations per person are available, one can estimate population mean and the population log-scale variance (V_1) for single measurements (M=1), and by averaging the two short-term measurements and then taking logs, one can estimate the population log-scale variance, V_2. (Sampling weights should be used when applicable.). By substituting into the above V_T equation for T=1 and T=2, the following formulas for estimating the variance components can be determined:

$$\hat{\gamma}_S = -\log[2\exp(\hat{V}_2 - \hat{V}_1) - 1]$$

and

$$\hat{\gamma}_L = \hat{V}_1 - \hat{\gamma}_S.$$

The distribution for any averaging time can then be estimated by choosing the appropriate M (e.g., M=365 if the measurement time is one day) and substituting estimates into the V_M equation above. Similarly, a "lifetime" distribution (also assumed to be lognormal) is then estimated by letting M go to infinity (i.e., the influence of the short term component vanishes). Wallace *et al.*(1994) caution that the data collection period should encompass all major long-term trends such as seasonality.

Clayton *et al.* (1998) describe a study of personal exposures to airborne contaminants that employs a more sophisticated study design and model (that requires more data); the goal was to estimate distributions of annual exposures from 3-day exposure measurements collected throughout a 12-month period. Two measurements per person (in different months) were available for some of the study participants. A multivariate lognormal distribution was assumed; the lognormal parameters for each month's data were estimated, along with the correlations for each monthly lag (assumed to depend only on the length of the lag). Simulated data were generated from this multivariate distribution for a large number of "people;" each "person's" exposures were then averaged over the 12 months. This approach assumes that the an average over 12 observations, one per month, produces an adequate approximation to the annual distribution of exposures. The model results were compared to those obtained via a modification of the DW model.

Buck *et al.* (1995, 1997) describe some general models (e.g., lognormally is not assumed); these, too, require multiple observations per person, and if the within-person variance is presumed to vary by person, then a fairly large number of observations per person may be needed. These papers give some insight into how estimated distributional parameters based on the short-term data relate to the long-term parameters. Reports by Carriquiry *et al.* (1995, 1996), Carriquiry (1996), and a paper by Nusser *et al.* (1996) deal with the some of the same issues in the context of estimating distributions of "usual" food intake and nutrition from short-term dietary data.

The second part of the inference — extrapolation from study time period (of duration τ) to the longer time T — is likely to be much less defensible than the first part, if τ and T are very different. This part of the inference is really an issue of temporal coverage. If the study involves person-day measurements conducted over a two-month period in the summer, and annual or lifetime inferences are desired, then little can be said regarding the relative variability or mean levels of the short-term and T-term data, basically because of uncertainty regarding the stationarity of the exposure factor over seasons and years. The above-described approach of Wallace *et al.*, for instance, includes statements that recognize the need for a population stationarity assumption that essentially requires that the processes underlying the exposure factor data that occur outside the time period of the surrogate population be like those that occur within the surrogate population. Applying some of the above methods on an age-cohort-specific basis, and then combining the results over cohorts, offers one possible way of improving the inference (e.g., see Hartwell *et al.*, 1992).

6. SUMMARY AND CONCLUSIONS

Representativeness is concerned with the degree to which "good" inferences can made from a set of exposure factor data to the population of concern. Thus evaluating representativeness of exposure factor data involves achieving an understanding of the source study, making an appraisal of the appropriateness of its internal inferences, assessing how and how much the surrogate population and population of concern differ, and evaluating the importance of the differences. Clearly, this can be an extremely difficult and subjective task. It is, however, very important, and sensitivity analyses should be included in the risk assessment that reflect the uncertainties of the process.

In an attempt to ensure that all aspects of representativeness are considered by analysts, we have partitioned the overall inferential process into components, some of which are concerned with design and measurement features of the source study that affect the internal inferences, and some of which are concerned with the differences between the surrogate population and the population of concern, which affect the external portion of the inference. We also partition the inferential process along the lines of the population characteristics — individual, spacial, and temporal — in an attempt to assess where overlaps and gaps exist between the data and the population of concern. In the individual and spatial characteristics, representativeness involves consideration of bounds and coverage issues. In the temporal characteristic, these same issues (i.e., study duration and currency) are important, but the time unit associated with the measurements or observations is also important, since time unit differences often occur between the data and the population of concern. Checklists are provided to aid in assessing the various components of representativeness.

When some aspect of representativeness is lacking in the available data, assessors are faced with the task of trying to make the data "more representative." We describe several techniques (and cite some others) for accomplishing these types of tasks; generally making such adjustments for known differences will reduce bias. However, it should be emphasized that these adjustment techniques cannot guarantee representativeness in the resultant statistics. For supporting future, large-scale (e.g., regional or national) risk assessments, one of the best avenues for improving the exposure factors data would be to get assessors involved in the design process -- so that appropriate modifications to the survey designs of future source studies can be considered. For example, the design might be altered to provide better coverage of certain segments of the population that may be the focus of risk assessments (e.g., more data on children could be sought). The use of multiple observations per person also could lead to improvement in those assessments concerned with chronic exposures.

7. BIBLIOGRAPHY

American Society for Quality Control (1994). *American National Standard: Specifications and Guidelines for Quality Systems for Environmental Data and Environmental Technology Programs (ANSI/ASQC E4)*. Milwaukee, WI.

Barton, M., A. Clayton, K. Johnson, R. Whitmore (1996). "G-5 Representativeness." Research Triangle Institute Report (Project 91U-6342-116), prepared for U.S. EPA under Contract No. 68D40091.

Buck, R.J., K.A. Hammerstrom, and P.B. Ryan (1995). "Estimating Long-Term Exposures from Short-Term Measuremetns." *Journal of Exposure Analysis and Environmental Epidemiology*, Vol. 5, No. 3, pp. 359-373.

Burmaster, D.E. and A.M. Wilson (1996). "An Introduction to Second-Order Random Variables in Human Health Risk Assessments." *Human and Ecological Risk Assessment*, Vol. 2, No. 4, pp. 892-919.

Carriquiry, A.L. (1996). "Assessing the Adequacy of Diets: A Brief Commentary" (Report prepared under Cooperative Agreement No. 58-3198-2-006, Agricultural Research Service, USDA, and Iowa State University).

Carriquiry, A.L., J.J. Goyeneche, and W.A. Fuller (1996). "Estimation of Bivariate Usual Intake Distributions" (Report prepared under Cooperative Agreement No. 58-3198-2-006, Agricultural Research Service, USDA, and Iowa State University).

Carriquiry, A.L., W.A. Fuller, J.J. Goyeneche, and H.H. Jensen (1995). "Estimated Correlations Among Days for the Combined 1989-91 CSFII" (Dietary Assessment Research Series Report 4 under Cooperative Agreement No. 58-3198-2-006, Agricultural Research Service, USDA, and Iowa State University).

Clayton, C.A., E.D. Pellizzari, C.E. Rodes, R.E. Mason, and L.L. Piper (1998). "Estimating Distributions of Long-Term Particulate Matter and Manganese Exposures for Residents of Toronto, Canada." Submitted to *Atmospheric Environment.*

Cohen, J.T., M.A. Lampson, and T.S. Bowers (1996). "The Use of Two-Stage Monte Carlo Simulation Techniques to Characterize Variability and Uncertainty in Risk Analysis." *Human and Ecological Risk Assessment*, Vol. 2, No. 4, pp. 939-971.

Corder, L.S., L. LaVange, M.A. Woodbury, and K.G. Manton (1990). "Longitudinal Weighting and Analysis Issues for Nationally Representative Data Sets." Proceedings of the *American Statistical Association, Section on Survey Research*, pp. 468-473.

Deville, J., Sarndal, C., and Sautory, O. (1993). "Generalized Raking Procedures in Survey Sampling." *Journal of the American Statistical Association*, Vol. 88, No. 423, pp. 1013-1020).

Ferson, Scott (1996). "What Monte Carlo Methods Cannot Do." *Human and Ecological Risk Assessment*, Vol. 2, No. 4, pp. 990-1007.

Folsom, R.E. (1991). "Exponential and Logistic Weight Adjustments for Sampling and Nonresurrogate populationonse Error Reduction." Proceedings of the *Social Statistics Section of the American Statistical Association*, 197-202.

Francis, Marcie and Paul Feder, Battelle Memorial Institute (1997). "Development of Long-Term and Short-Term Inhalation Rate Distributions." Prepared for Research Triangle Institute.

Hartwell, T.D., C.A. Clayton, and R.W. Whitmore (1992). "Field Studies of Human Exposure to Environmental Contaminants." Proceedings of the *American Statistical Association, Section on Statistics and the Environment*, pp. 20-29.

Holt, D. and Smith, T.M.F. (1979). "Post Stratification." *Journal of the Royal Statistical Society*, Vol. 142, Part 1, pp. 33-46.

Kendall, M.G. and W.R. Buckland (1971). A Dictionary of Statistical Terms. Published for the International Statistical Institute, Third Edition, New York: Hafner Publishing Company, Inc., p. 129.

Kruskal, W. and F. Mosteller (1979). "Representative Sampling, I: Non-Scientific Literature." *International Statistical Review*, Vol. 47, pp. 13-24.

Kruskal, W. and F. Mosteller (1979). "Representative Sampling, II: Scientific Literature, Excluding Statistics." *International Statistical Review*, Vol. 47, pp. 111-127.

Kruskal, W. and F. Mosteller (1979). "Representative Sampling, III: The Current Statistical Literature." *International Statistical Review*, Vol. 47, pp. 245-265.

Nusser, S.M., A.L. Carriquiry, K.W. Dodd, and W.A. Fuller (1996). "A Semiparametric Transformation Approach to Estimating Usual Daily Intake Distributions." *Journal of the American Statistical Association*, Vol. 91, No. 436, pp. 1440-1449.

Oh, H.L. and Scheuren, F.J. (1983). "Weighting Adjustment for Unit Nonresponse." In: *Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliographies*, Madow, W.G., Olkin, I., and Rubin, D.B., eds., Academic Press, New York, NY, pp. 143-184.

Shah, B., R. Folsom, L. LaVange, S. Wheeless, K. Boyle, and R. Williams (1993). *Statistical Methods and Mathematical Algorithms Used in SUDAAN*. Research Triangle Institute, Research Triangle Park, NC.

Smith, F., K. Kulkarni, L.E. Myers, and M.J. Messner (1988). Evaluating and Presenting Quality Assurance Sampling Data. In Keith, L.H. (Ed.), Principles of Environmental Sampling, American Chemical Society, ACS Professional Reference Book, pp. 157-168.

Stanek, III, E.J. (1996). "Estimating Exposure Distributions: A Caution for Monte Carlo Risk Assessment." *Human and Ecological Assessment*, Vol. 2, No. 4, pp. 874-891.

Wallace, L.A., Naihua Duan, and Robert Ziegenfus (1994). "Can Long-Term Exposure Distributions Be Predicted from Short-Term Measurements?." *Risk Analysis*, Vol. 14, No. 1, pp. 75-85.

CHECKLIST I. ASSESSING INTERNAL REPRESENTATIVENESS: POPULATION SAMPLED VS. POPULATION OF CONCERN FOR THE SURROGATE STUDY

- What is the study population?
    - What are the individual characteristics (i.e., defined by demographic, socioeconomic factors, human behavior and other study design factors)?
    - What are the spatial characteristics?
    - What are the temporal characteristics?
    - What are units of observation (e.g., person-days or person-weeks)?
    - What, if any, are the population subgroups for which inferences were especially desired?

- Are valid statistical inferences to the study population possible?
    - Was the whole population sampled (i.e., a census was conducted) used?
    - If not was the sample design appropriate and adequate?
        - Was a probability sample used? If not, how reasonable does the method of sample selection appear to be?
        - Was the response rate satisfactory?
        - Was the sample size adequate for estimating central tendency measures?
        - Was the sample size adequate for estimating other types of parameters (e.g., upper percentiles)?
        - For what population or subpopulation size was the sample size adequate for estimating measures of central tendency?
        - For what population or subpopulation size was the sample size adequate for estimating other types of parameters (e.g., upper percentiles)?
        - What biases are known or suspected as a result of the design or implementation or the study? What is the direction of the bias?

- Does the study appear to have and use a valid measurement protocol?
        - What is the likelihood of Hawthorne effects? What impact might this have on bias or variability?
        - What are other sources of measurement errors (e.g., recall difficulties)? What impact might they have on bias or variability?

- Does the study design allow (model-based) inferences to other time units?
        - What model is most appropriate?
        - What assumptions are inherent to the model?

CHECKLIST II. ASSESSING EXTERNAL REPRESENTATIVENESS: SURROGATE POPULATION VS. EXPOSURE ASSESSOR'S POPULATION OF CONCERN – INDIVIDUAL CHARACTERISTICS

- How does the population of concern relate to surrogate study population in terms of the individuals' characteristics?
  - Case 1: Are the individuals in the two populations essentially the same?
  - Case 2: Are the individuals in the population of concern a subset of those in the study population? If so, is there adequate information available to allow for the analysis of the population of concern? (Note: If so [Case 2a], we can redefine the surrogate data to include only persons in the population of concern and then treat this case as Case 1.)
  - Case 3: Are the individuals in the surrogate study population a subset of those in the population of concern?
  - Case 4: Are two populations disjoint -- in terms of individual characteristics?

- How important is the difference in the two populations (population of concern and surrogate population) with regard to the individuals' characteristics? To what extent is the difference between the individuals of the two populations expected to affect the population parameters?
  - With respect to central tendency of the two populations?
  - With respect to the variability of the two populations?
  - With respect to the shape and/or upper percentiles of the two populations?

- Is there a reasonable way of adjusting or extrapolating from the surrogate population to the population of concern -- in terms of the individuals' characteristics?
  - What method(s) should be used?
  - Is there adequate information available to implement it?

CHECKLIST III. ASSESSING EXTERNAL REPRESENTATIVENESS: SURROGATE POPULATION VS. EXPOSURE ASSESSOR'S POPULATION OF CONCERN -- SPATIAL CHARACTERISTICS

- How does the population of concern relate to surrogate population in the spatial characteristics?
    - Case 1: Do they cover the same geographic area?
    - Case 2: Is the geographic area of the population of concern a subset of the area of surrogate population? If so, is there adequate information available to allow the analysis of the population of concern? (Note: If so [Case 2a], we can redefine the surrogate population to include only regions or types of geographic areas in the population of concern and then treat this case as Case 1.)
    - Case 3: Is the geographic area covered by the surrogate population a subset of that covered by the population of concern?
    - Case 4: Are two populations disjoint -- in the spatial characteristics?

- How important is the difference in the two target populations with regard to the spatial characteristics? To what extent is the difference in the spatial characteristics of the two populations expected to affect the population parameters?
    - With respect to central tendency of the two populations?
    - With respect to the variability of the two populations?
    - With respect to the shape and/or upper percentiles of the two populations?

- Is there a reasonable way of adjusting or extrapolating from the surrogate population to the population of concern -- in terms of the spatial characteristics?
    - What method(s) should be used?
    - Is there adequate information available to implement it?

CHECKLIST IV. ASSESSING EXTERNAL REPRESENTATIVENESS: SURROGATE POPULATION VS. EXPOSURE ASSESSOR'S POPULATION OF CONCERN -- TEMPORAL CHARACTERISTICS

- How does the population of concern relate to surrogate population in terms of currency and temporal coverage (study duration)?
  - Case 1: Are the duration and currency of the surrogate data compatible with the population of concern needs?
  - Case 2: Is the temporal coverage of the population of concern a subset of the surrogate population? If so, is there adequate information available to allow the analysis of the population of concern? (Note: If so [Case 2a], we can redefine the surrogate population to include only time periods (e.g., seasons) of interest to the assessor and then treat this case as Case 1.)
  - Case 3: Is the temporal coverage of the surrogate population a subset of that covered by the population of concern?
  - Case 4: Are the two populations disjoint — in terms of study duration and currency?

- How does the population of concern relate to surrogate population in terms of the time unit (either the observed time unit or, if appropriate, a modeled time unit)?
  - Case 1: Are the time units compatible?
  - Case 2: Is the time unit for the population of concern shorter than that of the surrogate population? If so, are data available for the shorter time unit associated with the population of concern. (If so [Case 2a], this can be treated as Case 1.)
  - Case 3: Is the time unit for the population of concern longer than that of the surrogate population?

- How important is the difference in the two populations (i.e., population of concern and surrogate population) with regard to the temporal coverage and currency? To what extent is the difference in the temporal coverage and currency of the two populations expected to affect the population parameters?
  - With respect to central tendency of the two populations?
  - With respect to the variability of the two populations?
  - With respect to the shape and/or upper percentiles of the two populations?

- Is there a reasonable way of adjusting or extrapolating from the surrogate population to the population of concern -- to account for differences in temporal coverage or currency?
  - What method(s) should be used?
  - Is there adequate information available to implement it?

- How important is the difference in the two populations (i.e., population of concern and surrogate population) with regard to the time unit of observation? To what extent is the difference in the observation time unit of the two populations expected to affect the population parameters?
  - With respect to central tendency of the two populations?
  - With respect to the variability of the two populations?
  - With respect to the shape and/or upper percentiles of the two populations?

- Is there a reasonable way of adjusting or extrapolating from the surrogate population to the population of concern -- to account for differences in observation time units?
  - What method(s) should be used?
  - Is there adequate information available to implement it?

Issue Paper on Empirical Distribution Functions and Non-parametric Simulation

Introduction

One of the issues facing risk assessors relates to the best use of empirical distribution functions (EDFs) to represent stochastic variability intrinsic to an exposure factor. Generally, one of two situations occurs. In the first situation, the risk assessor is reviewing an assessment in which an EDF has been used. The risk assessor needs to make a judgement whether or not the use of the EDF is appropriate for this particular analysis. In the second situation, the risk assessor is conducting his/her own assessment and must decide whether a parametric representation or non-parametric representation is best suited to the assessment. The objective of this issue paper is to help focus discussion on the key issues and choices facing the assessor under these circumstances.

We make the initial assumption that the data are sufficiently representative of the exposure factor in question. Here, representative is taken to mean that the data were obtained as a simple random sample of the relevant characteristic of the correct population, that the data were measured in the proper scale (time and space), and that the data are of acceptable quality (accuracy and precision).

We also make the assumption that the analysis involves an exposure/risk model which includes additional exposure factors, some of which also exhibit natural variation. Ultimately, we are interested in estimating some key aspects of the variation in predicted exposure/risk. As a minimum, we are interested in statistical measures of central tendency (e.g., median), the mean, and some measure of plausible upper bound or high-end exposure (e.g., 95th, 97.5th, or 99th percentiles of exposure). Thus, how variable factors algebraically and statistically interact is important.

Further, we assume that Monte Carlo methods will be used investigate the variation in exposure/risk. Obviously, other methods can be used, but it is clear from experience that simulation-based techniques will be used in the vast majority of applications.

Conventional wisdom advises that when there is an underlying theory supporting the use of a particular theoretical distribution function (TDF), then the data should be used to fit the distribution and that distribution should be used in the analysis. For example, it has been argued that repeated dilution and mixing of an environmental pollutant should eventually result in a lognormal distribution of concentrations. While this is an agreeable concept in principle, it is rare situation when a theory-based TDFs are available for particular exposure factors. Furthermore, theory-based TDFs are often only valid in the asymptotic sense. Convergence is may be very slow, and, in the early stages, the data may be very poorly modeled by the asymptotic form of the TDF. For this issue paper, we assume that no theory-based TDFs are available.

The issue paper is written in two parts. Part I addresses the strengths and weakness of empirical distribution functions; Part II addresses issues related to judging quality of fit for theoretical distributions.

Part I. Empirical Distribution Functions

Definitions. Given representative data, X = {$x_1, x_2, \cdots, x_n$ }, the risk assessor has two basic techniques for representing an exposure factor in a Monte Carlo analysis:

parametric methods which attempt to characterize the exposure factor using a TDF. For example, a lognormal, gamma, or Weibull distribution is used to represent the exposure factor, and the data are used to estimate values for its intrinsic parameters.

non-parametric methods which use the sample data to define an empirical distribution function (EDF) or modified version of the EDF.

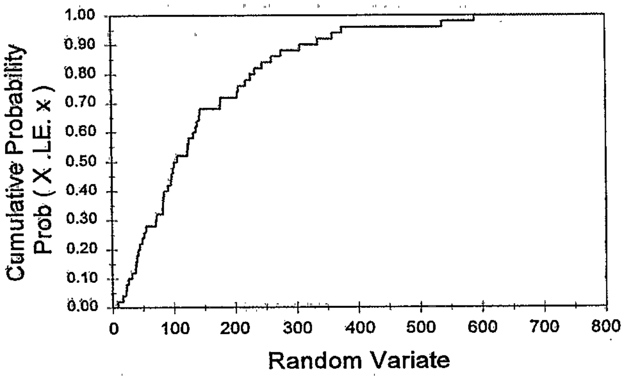
EDF. Sorted from smallest to largest, $x_1 \le x_2 \le \cdots x_n$, the EDF is the cumulative distribution function defined by

$$\hat{F}(x) = \frac{number\ \ of\ x_k \le x}{n} \qquad or \qquad \hat{F}(x) = \frac{1}{n}\sum_{k=1}^{n} H(x - x_k)$$

where H(u) is the unit step function which jumps from 0 to 1 when u ≥ 0. The values of the EDF are the discrete set of cumulative probabilities (0, 1/*n*, 2/*n*, ···, *n*/*n*). Figure 1 illustrates a basic EDF for 50 samples drawn from lognormal distribution with a geometric mean of 100 and a geometric standard deviation of 3, i.e., X ~ LN(100,3).

In a Monte Carlo simulation, an EDF is generated by randomly sampling the raw data with replacement (simple bootstrapping) so that each observation in the data set, x_k, has an equal probability of selection, i.e., $prob(x_k) = 1/n$.

Figure 1. Example of EDF



Properties of the EDF. The following summarizes some of the basic properties of the EDF:

1. Values between any two consecutive samples, x_k and x_{k+1} cannot be simulated, nor can values smaller than the sample minimum, x_1, or larger than the sample maximum, x_n, be generated, i.e., $x \geq x_1$ and $x \leq x_n$

2. The mean of the EDF is equal to the sample mean. The variance of the EDF mean is always smaller than the variance of the sample mean; it is equal to $(n-1)/n$ times the variance of the sample mean.

3. The variance of the EDF is equal to $(n-1)/n$ times the sample variance.

4. Expected values of the EDF percentiles are equal to the sample percentiles.

5. If the underlying distribution is skewed to the right (as are many environmental quantities), the EDF will tend to under-estimate the true mean and variance.

Figures 2 and 3 below illustrate typical Monte Carlo behavior of the EDF in reproducing the sample mean, variance, and 95th percentile of the underlying sample. Here X ~ LN(100,3) with a sample size of $N = 100$ and the relative error is defined as 100× [simulated−sample]/sample. The oscillatory nature of the simulated 95th percentile reflects the normalized magnitude of the difference between adjacent order statistics in the sample, $x_{(95)}$, and $x_{(96)}$ and shows the Monte Carlo estimate flip-flopping between these two ranks

Figure 2. Convergence of the Mean and Varianc

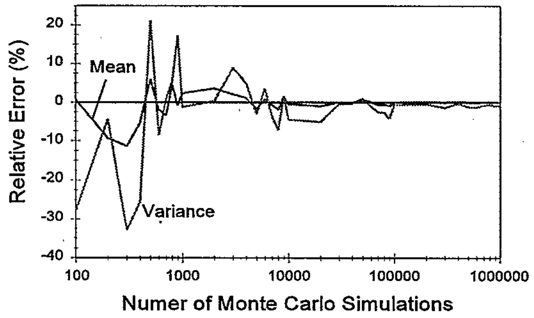
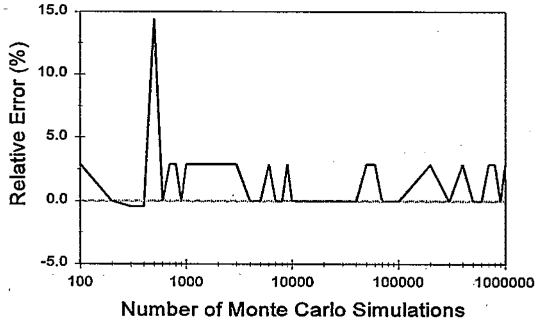


Figure 3. Convergence of the 95th Percentile

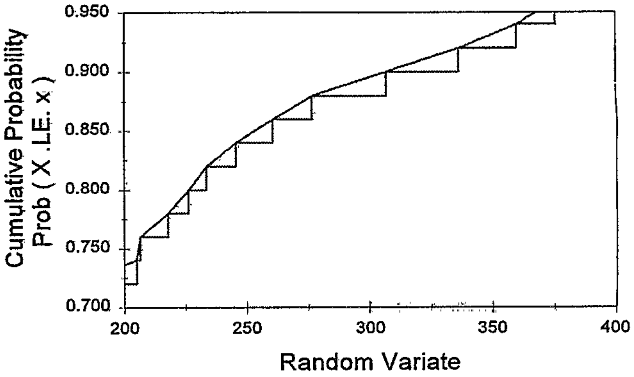


Linearly Interpolated EDF (Linearized EDF). For continuous random variables, it may be troubling to define the EDF as a step function and so extrapolation is often used to estimate the probabilities of values in between sample values. Generally, for values between observations, linear interpolation is favored, although higher order interpolation is sometimes used. Figure 4 compares a linearly interpolated EDF with the basic EDF. The linearly interpolated EDF will tend to underestimate the sample mean and variance. It will converge to the appropriate sample percentile, but take longer to do so when compared to the simple EDF. These differences tend to diminish as the sample size increases. Table 1 illustrates differences between the EDF, linearized EDF and best fit TDF for residential room air exchange rates. The EDF statistics are based on a Monte Carlo simulation with 25,000 replications. Clearly the simple EDF is best at reproducing sample moments and sample percentiles.

Statistic	ACH Sample N = 90	EDF	Linearized EDF	Best Fit Weibull PDF
mean	0.6822	0.6821	0.6747	0.6782
variance	0.2387	0.2358	0.2089	0.2479
skewness	1.4638	1.4890	1.2426	1.2329
kurtosis	6.6290	6.7845	5.6966	4.9668
5%	0.1334	0.1320	0.1307	0.0881
10%	0.1839	0.1840	0.1840	0.1452
50%	0.6020	0.6160	0.6032	0.5691
90%	1.2423	1.2390	1.2398	1.3592
95%	1.3556	1.3820	1.3600	1.6450

Table 1 Comparison of key summary statistics

Figure 4. Comparison of Basic EDF and Linearly Interpolated EDF



Extended EDF. Neither the simple EDF nor the interpolated EDF can produce values beyond the sample minimum or maximum. This may be an unreasonable restriction in many cases. For example, the probability that a previously observed largest value in a sample based on n observations will be exceeded in a sample of N future observations may be estimated using the relationship $prob = 1 - n/(N + n)$. If the next sample size is the same as the original sample size, there is a 50% likelihood that the new sample will have a largest value greater than the original sample's largest value. Restricting the EDF to the smallest and largest sample values will produce distributional tails that are too short. In order to get around this problem, one may extend the EDF by adding plausible lower and upper bound values to the data. The actual values are usually based on theoretical considerations or on expert judgement. For right skewed data, adding a new minimum and maximum would tend to increase the mean and variance of the EDF. This same sort or rational is used when continuous, unbounded TDFs are truncated at the low and high end to avoid generating unrealistic values during Monte Carlo simulation (e.g., 15 kg adult males, females over 2.5m tall, etc.)

Mixed Empirical-Exponential Distribution. An alternative approach to extending the upper tail of an empirical distribution beyond the sample data has been suggested by Bratley *et al.* In their method, an exponential tail is fit to the last five or ten percent of the data. This method is based on extreme value theory and the observation that extreme values for many continuous, unbounded distributions follow an exponential distribution.

Starting Points

The following table summarizes the results of an informal survey of experts who were asked to contribute their observations and thoughts on the strengths and weaknesses of EDFs by addressing a list of questions and issues. Based on this survey:

1. The World seems to be divided into TDF'ers and EDF'ers.

2. There are no clear-cut, unambiguous statistical reasons for choosing EDFs over TDFs or vice versa.

3. Many of the criticisms leveled at EDFs also apply to TDFs (e.g., the data must be simple random samples)..

4. One aspect of which may have important implications for our discussion is the nature of the decision and how sensitive an outcome is to the choice of an EDF.

5. Generally, contributors did not express much support for either the linearized EDF or the extended EDF. Why they seem to be comfortable with TDFs, which essentially interpolate between data points as well as extrapolated beyond the data, is unclear.

Issue	Comments
1. EDFs provide complete representation of the data without any loss of information.	Yes, but perhaps an incomplete representation of what is known about the quantity for which the distribution is needed.
2. EDFs do not depend on any assumptions associated with parametric models.	One has to assume a representative random sample. As another example, advantage 2 (EDFs do not depend on parametric assumptions) is true and is a well-known advantage. Less well known is that almost all non-parametric procedures make some strong assumptions. Technically, a parametric situation is one where you limit the class of possible probability distributions to a collection that can be described in a natural way using a finite number of real numbers, or parameters. In common non-parametric situations (such as comparing medians of two sets of data) the data are modeled by pairs of distributions, but there is still a restriction, such as that the members of each pair are the same distribution except for a change of location. Furthermore, using an EDF is something entirely different than the set of assumptions you make about the class of possible distributions. Usually, you use an EDF as a tool to make an estimate: that is, as a computational device.
3. For large samples, EDFs converge to the true distribution for all values of X.	Although for most well-behaved distributions it is the case that the EDF converges in probability to the underlying distribution, convergence often requires unrealistic amounts of data. One important issue in risk assessment is the near universal situation of having too few data. This usually means we are nowhere near a limiting case and that we should beware ALL asymptotic methods, including Maximum Likelihood, without careful evaluation of their applicability to our small data set. EDFs usually converge VERY slowly to the underlying distribution (especially if you're trying to characterize extreme events). Therefore this convergence phenomenon is not very comforting or useful. EDFs are almost useless, except in very large data sets. Accuracy of any interval is driven by a standard deviation of sqrt(n) in that interval. For even a 10% accuracy, with 20 intervals, you would need more than 2,000 underlying observations. This is useless, since "large" is unattainable for all practical purposes, unless you're the Census Bureau.
4. EDFs provide direct information on the shape of the underlying distribution, e.g., skewness and bimodality; EDFs supply robust information on location and dispersion.	Yes, but the confidence limits on those estimates can be quite wide in some cases. For example, a small data set that is negatively skewed could be a random sample from a positively skewed population.
5. An EDF can be an effective indicator of peculiarities (e.g., outliers)	Maybe. Not sure how this is different than when comparing data to a fitted parametric distribution or mixture distributions.
6. An EDF does not involve grouping difficulties and loss of information associated with the use of histograms	True.

7. Confidence intervals are easily calculated.	For what? how? They can be calculated or simulated for parametric distributions as well. Not sure why this is an advantage for EDFs and not parametric distributions also. It's nice when confidence intervals are easily calculated, but usually the more important criteria are whether they have the coverage claimed of them and how tight the intervals are. Yes, but crude if measurements are limited. The biggest advantage of EDFs you left out: free from subjective model bias. I.e., the choice of parametric form may affect conclusions.
8. EDFs can be sensitive to random occurrences in the data and sole reliance on them can lead to spurious conclusions. This can be especially true if the sample size small	This is true in all cases with small data sets. The best thing is to consider confidence intervals on the distributions to get an idea of whether the occurrences might be random or real. This is ONLY true if the sample size is small. This is the very essence of the issue.

9. How much data do I need to develop a useful EDF?	What you need is random representative data and to feel comfortable that your data include the lower and upper bounds of the quantity. The number of data points in itself is not particularly important. How many data? Two. This somewhat flippant answer simply highlights the important fact that you need to ask the question in the context of (a) what decision is being made and (b) what its risk function is (how bad is it if the decision is incorrect?). If the risk function is low (it doesn't matter much if we are wrong) and the decision is really obvious, then sometimes all you need is a reality check. Hence the need for one datum. People make mistakes and Murphy's Law applies, so experience dictates a second datum. I know you guys at EPA and in the states are competent and sensible and often very good at this stuff, but there are still many people and many agencies out there that are just too uncomfortable with common sense like this, so it pays to repeat it. (The comment cuts both ways: sometimes I am asked by clients to gather more data to show that they don't have a problem, when all their data point to serious contamination. Most of them back down right away when confronted with the common-sense approach--"you obviously have a problem, so let's talk instead about how to remedy it, since honest statistics won't make it go away.") I would not approach the topic this way. I would ask, instead, how do I characterize an amount of data, and given these summary characteristics, what methods are appropriate. At a minimum 10 points per interval needed, with about 10-20 intervals usually needed for reasonable interpolation of most density curves. For bimodal, etc., double the number of intervals. Gee, that depends. I think the main consideration is the importance of the tails in the decision. If you are going to place a lot of weight on the 99th percentile, then 100 data points are telling you want you want to know. If you are primarily interested in the average or 90th percentile, then 100 data points is pretty good. This is similar to the "how many iterations is enough?" problem. If you have as many data points as iterations, then I think it is pretty hard to justify NOT using an EDF If you are going to place a lot of weight on the 99th percentile, then 100 data points are telling you want you want to know. EXCEPTION: Not with much accuracy. The theory is simple and one example will illustrate the issue. By definition of percentile, there is 0.99 probability that a value above the 99th percentile of the (true) underlying distribution does not occur in a random sample. In a sample of 100 data selected independently from that distribution, values between the 99th and 100th percentiles therefore do not occur with probability (0.99)^100, which is extremely close to 1/e, or almost 40%. Therefore there are almost even odds (2:3) that with 100 data you have not even seen anything as high as the 99th percentile yet. To be fairly sure of seeing a value that high, you need to solve (0.99)^N <= Assurance value (such as 5%) for N. That would require about N=300 points in this example, and even then you only have 95% confidence that you have seen A SINGLE value at or above the 99th percentile.

10. Should I linearize the EDF between percentiles or use step functions?	A true EDF uses step functions--this is resampling of the data in which each data point has a probability 1/n. The use of linear interpolation will typically lead to lower estimates of the standard deviation, since you are not guaranteed to sample the min and max data points. Now you're going down a slippery slope. As soon as you linearize your EDF you are entering into the land of semi-parametric techniques, smoothing, modeling, and assumptions. You're not using the EDF any more. The EDF is accurately and correctly described by its cumulative distribution function, which will be a step function. If your aren't using a continuous distribution, why not just go with the data? The diversity of distributions is very rich. For example, see Evans, Hasting, and Peacock, Statistical Distributions, 2nd Ed., Wiley (1993) for 39 of them. Using some kind of test for fit of the continuous distribution to your data, e.g., quantiles, you usually can obtain a reasonable fit. See JW Tukey, Exploratory Data Analysis Addison-Wesley (1977). If not, e.g., bimodal, you will have to decompose or transform your data, and you already start to make important assumptions. Smoothing EDFs within the bulk of the probability curve causes no serious errors. Extrapolation beyond the limits of data violates the very concept of EDF, and is intrinsically dependent on the parameterization used. The simple solution is to use the midpoint rule (apply prob. at the interval midpoint). Alternatively, use trapezoidal rule (st. line interpolation). For a continuous curve, a straight line interpolation averages properly and improves discretization bias. I, however, would suggest using resampling as a better approach than smoothing. I usually use percentiles, but you have enough data to use an EDF, then it shouldn't matter much.
11. When the data set is large, should I bin the data into a histogram to speed up the simulation? If so, what defines large? How does it depend on the distribution of the data?	In this case, the difference between step functions and linear interpolations becomes small. Why bin? You lose information that way. If you have large segments of the CDF that are approximately piecewise uniform, then binning the data won't result in much loss of information. here's a lot of literature on binning data, mostly in terms of how the perception of the histogram can change. I would suggest, in the spirit of the response to question 1, that you consider the effect the binning process has on the outcome of your work, since your question really is one of computational practice, not conceptual approach. Bin the data to speed your process (simulation, bootstrapping, whatever) but in a way in which you can demonstrate your answers are not materially different than what you would get with a more accurate procedure. How do you know what a material difference is? Look at your decision space and your risk function. No! This approach causes more mischief in epidemiology than in exposure analysis, but anytime you summarize the data, you lose information. If the data set is large, feel grateful. The intervals or bins used are mathematical estimators of the underlying density or distribution curve. This is a numerical integration or interpolation issue. Typically 10-20 intervals gives good performance on a unimodal density function. Particularly if linear interpolation is used. No.

12. Should I add a minimum and maximum to the data set so that points outside the observed data can be generated during simulation? The min, max could be based on theoretical considerations or expert judgment.	Why not just use an appropriate parametric distribution instead. This is where the "empirical" approaches fall flat on their face. Some have proposed these bizarre mixed empirical-exponential distributions with exponential and polynomial extrapolations based upon the largest and smallest data points... this can't be defended other than as arbitrary. In contrast, there may be a mechanistic basis for selecting a parametric distribution. You're sliding further down the slope. Adding a min or max and using theory or expert judgment seemed to be just what you wanted to avoid by using an EDF. If you're going to do that, you're wide open to criticism. Perhaps better to use some of the other procedures you mention, such as exponential tail fitting. However, if these kinds of procedures will not really change the answer in a material way, go for it. Again, no. Let the data talk to you. I punt. This is a tail problem that arises when the data really isn't telling you what you want to know. Whatever you base you judgment on will have to be based on other evidence.
13. Should I consider a mixed empirical-exponential distribution? An method for extending the upper tail of an EDF beyond the sample data has been suggested by Bratley et al. In their method, an exponential tail is fit to the last five or ten percent of the data. This method is based on extreme value theory and the observation that extreme values for many continuous, unbounded distributions follow an exponential distribution. The exponential tail is fit so that the mean of the data set is conserved.	I don't like this method as described above. I don't see what it offers in contrast to parametric distributions, and it would seem to open the analyst up for excessive criticism. I like the mixed distribution approach (after having carefully read Gnedenko's original paper on extreme value distributions to understand how applicable this approach is). Often you can produce good theoretical and statistical reasons why the tail of your data represents a random sample of extreme values. You need to have this justification, though, since not all probability tails are exponential, and some are very far from it. It's no problem to do this, and it may be fun to see what you get, but any conclusions you reach depend entirely on the assumptions in method and your fitting process. You could also (and I would somewhat prefer) using more complex (e.g., biphasic) distribution functions that allow more freedom to fit tail data.

14. If I bootstrap and if the exposure variable is continuous, what should I do, if anything, about values in between my data points which will not be simulated in the resampling process?	Probably nothing needs to be done if you assume the data are a representative random sample. The answer will look noisy or jumpy due to the gaps in the data, but that in and of itself is not a bad thing. Use of linear interpolations can lead to different estimates of standard deviations and other statistics when compared with the step-wise EDF. You have partially answered this question with the exponential tail fitting suggestion. When you start interpolating and fitting curves to your EDF, you are no longer in the purely parametric realm and you forgo a lot of the EDF advantages you so carefully listed--but sometimes you can't trust using just the EDF. As I understand bootstrap, you must generate a distribution, by parameterizing your data, as a first step. This step takes care of interpolation. You could bootstrap from percentiles, or take percentiles from bootstraps. I wouldn't think it would make much difference.

Part II. Issues Related to Fitting Theoretical Distributions

Suppose the following set of circumstances:

(1) that we have a random sample of an exposure parameter which exhibits natural variation

(2) that the collected data are representative of the exposure parameter of interest (i.e., the data measure the right population, in the right time and spatial scales etc.)

(3) that estimates of measurement error are available.

(4) that there is no available physical model to describe the distribution of the data (i.e., there is no theoretical basis to say that the data are lognormal, gamma, Weibull, etc).

(5) that we wish to characterize and account for the variation in the parameter in an analysis of environmental exposures.

(6) we run the data through our favorite distribution-fitting software and get goodness of fit statistics (e.g., chi-square, Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling, Watson, etc.) and their statistical significance.

(7) rankings based on the goodness of fit results are mixed, depending on the statistic and p-values.

(8) graphical examination of the quality of fit (QQ plots, PP plots, histogram overlays, residual plots, etc) presents a mixed picture, reinforcing the differences observed in the goodness of fit statistics.

Questions

1). A statistician might say that one should pick the simplest distribution not rejected by the data. But what does that mean when rejection is dependent on the statistic chosen and an arbitrary level of statistical significance?

2). On what basis should it be decided whether or not a data set is adequately represented by a fitted analytic distribution?

3). Specifically, what role should the p-value of the goodness of fit statistic play in that judgment?

4). What role should graphical examination of fit play?

Respondent #1
All distributions are, in fact empirical. Parametric distributions are merely theoretical constructs. There is no reason to believe that any given distribution is, in fact, log-normal (or any other specific parametric type). That we agree to call a distribution log-normal is (or at least should be) merely a shorthand by which we mean that it looks sufficiently like a theoretical log-normal distribution to save ourselves the extra work involved in specifying the empirical distribution. Other than analyses where we are dealing strictly with hypothetical constructs (e.g, what if we say that such-and-such distribution is lognormal and such and such distribution is normal....), I can see no theoretical justification for a parametric distribution other than the convenience gained. When the empirical data are sparse in the tails, we, of course, run into trouble in needing to specify an arbitrary maximum and minimum to the empirical distribution. While this may introduce considerable uncertainty, it is not necessarily a more uncertain practice than allowing the parametric construct to dictate the shape of the tails, or for that matter arbitrarily truncating the upper tail of a parametric distribution. This becomes less of a problem if the analysts goal in constructing an input distribution is to describe the existing data with as little extrapolation as necessary rather than to predict the "theoretical" underlying distribution. This distinction gets us close to the frequentist/subjectivist schism where many, if not all MC roads eventually seem to lead.

Respondent #2
...if you use p-bounds you don't have to choose a single distribution. You can use the entire equivalence class of distributions (be it a large or small class). I mean, if you can't discriminate between them on the basis of goodness of fit, maybe you do the problem a disservice to try. And operationalizing the criterion for "simplest" distribution is no picnic either.

Respondent #3
Why not try the KISS method: Keep It Simple & Sound. The Ranked Order Data assuming uniform probability intervals is a method that makes no assumptions as to the nature of the distribution. I also tends to the true distribution function as the number of data points increases. If you have replicate measurements (on each random sample) then the mean of these should be used.

The method yields simple rapid random number generators and one can obtain and desired statistical parameter of the distribution. However, use of the distribution function in any estimate is advised. Given the high level of approximation and/or bias in most risk assessment data and models, any approximation to the true PDF should be adequate.

There is one occasion when the theoretical PDF may be better than the empirical PDF. That is when it comes from the solution of equations based on fundamental laws constraining the solution to a specified form. Even in this case agreement with data is required. This in not usually the case in risk assessment PDFs.

Respondent #4
Since I am blessed not to be a statistician, I have no problem disputing their "statement" about the "simplest" distribution. I don't know what they mean either. What really matters physically is picking a distribution that has the fewest variables and that is easy to apply, given the kind of analysis you want to do. You want one that does not make assumptions in its construction that contradict processes operating in your data. If your are generating equally bad fits with a variety of the usual distributions anyway, by all means chose the one that is easiest to use. For time sliced exposure data, the "right" distribution almost always means a lognormal distribution. A physical basis for the lognormal does exist for exposure data, and empirically, most exposure data fit lognormals. [Your assumption "A" does not hold for typical exposure processes.] Wayne Ott, who probably does not even remember it, taught me this one afternoon in the back of a meeting room. See "A Probabilistic methodology for analyzing water quality effects of urban runoff on rivers and streams," Office of Water, February 15, 1984. Just tell people that you have used a lognormal distribution for convenience, although it does not fit particularly well, then provide some summary statistics that describe the poorness of fit.

Problems begin when you get a poor fit to a lognormal distribution but a good fit with a different distribution. Say you get a better fit to the Cauchy distribution, because the tails of your pdf have more density. Now things get more fun. Statisticians would say that you should use the Cauchy distribution, because it is a better fit. I say that you should still use the lognormal, because you can interpret manipulations of the data more easily, and just note that the lognormal fit is poor. Problems will arise, however, if you want to reach conclusions that rely on the tails of the distribution, and you use the lognormal pdf formulation, instead of your actual data. I somewhat anticipated your dilemma in my previous E-mail to you. If you don't need to use a continuous distribution, just go with the data!"

For time dependent exposure data, the situation gets much more complex. I prefer to work with Weibull distributions, but I see lots of studies that use Box-Jenkins models.

And you also asked: On what basis do I decide whether my data are adequately represented by a fitted analytic distribution? Specifically, what role should the p-value of the goodness of fit statistic play in my choice? What role should graphical examination of fit play?

To me, the data are adequately represented, when the analytical distribution adequately fills the role you intend it to have. In other words, if you substitute a lognormal distribution for your data, as a surrogate, then carry out some operations and obtain a result, the lognormal is adequate, unless it leads to a different conclusion than the actual data would support. The same statement is true of any continuous distribution.

Similarly, as a Bayesian, I think that the proper role of a p-value is the role you believe it should play. I don't think that p-values have much meaning in these kinds of analyses, but if you think they should, you should state the desired value before beginning to analyze the data, and not proceed until you obtain this degree of fittedness or better. If small differences in p-value make much difference in your analysis, your conclusions are probably too evanescent to have much usefulness. The quantiles approach that I previously commended to you, is a graphical method. [See J.W. Tukey, Exploratory Data Analysis. Addison-Wesley (1977)]. In it, you would display the distribution of your data, mapped against the prediction from the continuous distribution you have chosen, with both displayed as order statistics. If your data fit your distribution well, the points (data quantiles versus distribution quantiles, will fall along a straight (x=y) line. Systematic differences in location, spread, and/or shape will show up fairly dramatically. Such visual inspection is much more informative than perusing summary statistics. No "statistical fitting" is involved. [Also see J.M. Chambers et al., Graphical Methods for Data Analysis. Cole Publishing (1983)].

Respondent #5

I have several thoughts on the goodness of fit question. First, visual examination of the data is likely to yield more insight into the REASONS for the mixed behavior of the various statistics; i.e., in what regions of the variable of interest does a particular theoretical distribution not fit well, and in what direction is the error? Then choosing a particular parametric distribution can be influenced by the purpose of the analysis. For example, if you are interested in tail probabilities, then fitting well in the tails will be more important than fitting well in the central region of the distribution, and vice versa.

A good understanding of the theoretical properties of the various distributions is also handy. For example, the heavy tails of the lognormal mean that the moments can be very strongly influenced by relatively low-probability tails. If that seems appropriate fine; if not the analyst should be aware of that, etc. I don't think there is a simple answer; it all depends on what you are trying to do and why!

Respondent #6

In broad overview, I have these suggestions -- all of which are subject to modification, depending on the situation.

1. Professional judgment is **unavoidable** and is **always** a major part of every statistical analysis and/or risk assessment. Even a (dumb) decision to rely **exclusively** on one particular GOF statistic is an act of professional judgment. There is no way to make any decision based exclusively on "objective information" because the decision on what is considered objective contains unavoidable subjective components. There is no way out of any problem except to use and to celebrate professional judgment. As a profession, we risk assessors need to get over this hang up and move ahead.

2. It is **always** necessary and appropriate to fit several different parametric distributions to a data set. We make choices on the adequacy of a fit by comparison to alternatives. Sometimes we decide that one 2-parameter distribution fits well enough (and better than the reasonable

alternatives) so that we will use this distribution. Sometimes we decide that it is necessary to use a more complicated parametric distribution (e.g., a 5-parameter "mixture" distribution) to fit the data well (and better than the reasonable alternatives). And sometimes, we decide that no parametric distribution can do the job adequately well, hence the need for bootstrapping and other methods.

3. The human eye is far, far better at **judging** the overall match (or lack thereof) between a fitted distribution and the data under analysis than any statistical test ever devised. GOF tests are "blind" to the data! We need to visualize, visualize, and visualize the data -- as compared to the alternative fitted distributions -- to **see** how the various fits compare to the data. Mosteller, Tukey, and Cleveland, three of the most distinguished statisticians of the last 50 years, have all stressed the **essential** nature of visualization and human judgment relying thereon (in lieu of GOF tests). BTW, these graphs and visualizations *must* be published for all to see and understand.

4. In situations where no single parametric distribution provides an **adequate** fit to the data, there are several possible approaches to keep moving ahead. Here are my favorites.

A. (standard approach) Fit a "mixture" distribution to the data.

B. Use the two or three or four parametric distributions that offer the most appealing fit in a sensitivity analysis to see if the differences among the candidate distributions really make a difference in the decision at hand. Get the computer to simulate the results of choosing among the different candidate distributions. This leads to keen insights as to the "value of information".

C. (see references below, and references cited therein) By extension of the previous idea, analysts can fit and use "second-order" distributions that contain both **Variability** and **Uncertainty**. These second-order distributions have many appealing properties, especially the property that they allow the analyst to propagate Variability and Uncertainty **separately** so the risk assessor, the risk manager, and the public can all see how the Var and Unc combine throughout the computation / simulation into the final answer.

Respondent #7
[RE comments #1, #3, respondent #6]. ... the motivation behind having standardized methods: Professional judgment does not always produce the same result. Your professional judgment does not necessarily coincide with someone else's professional judgment. Surely, you've noticed this. The problem isn't that no one is celebrating their professional judgement - the problem is that we have more than one party.

The bigger and more unique the problem, the less standardization matters. But if you are trying to compare, say, the risk from thousands of superfund sites, you can't very well reinvent risk analysis for every one and expect to get comparable results - whatever you do for one you must do for all.

Have you tried to produce a GOF statistic that matches your visual preference? I have. For instance, I think fitting predicted percentiles produces better looking fits than fitting observed values (e.g., maximum likelihood) - because this naturally gives deviations at extreme values less weight - where 'extreme value' is model dependent.

APPENDIX B

LIST OF EXPERTS AND OBSERVERS

EPA United States
Environmental Protection Agency
Risk Assessment Forum

Workshop on Selecting Input Distributions for Probabilistic Assessment

U.S. Environmental Protection Agency
New York, NY
April 21-22, 1998

List of Experts

Sheila Abraham
Environmental Specialist
Risk Assessment/Management
Northeast District Office
Ohio Environmental
Protection Agency
2110 East Aurora Road
Twinsburg, OH 44087
330-963-1290
Fax: 330-487-0769
E-mail:sabraham@epa.state.oh.us

Hans Allender
U.S. Environmental
Protection Agency
401 M Street, SW (7509)
Washington, DC 20460
703-305-7883
E-mail: allender.hans@
epamail.epa.gov

Timothy Barry
Office of Science Policy,
Planning, and Evaluation
U.S. Environmental
Protection Agency
401 M Street, SW (2174)
Washington, DC 20460
202-260-2038
E-mail: barry.timothy@
epamail.epa.gov

Robert Blaisdell
Associate Toxicologist
California Office of Environmental
Health Hazard Assessment
2151 Berkeley Way
Annex 11 - 2nd Floor
Berkeley, CA 94704
510-540-3487
Fax: 510-540-2923
E-mail: bblaisde@
berkeley.cahwnet.gov

David Burmaster
President
Alceon Corporation
P.O. Box 382669
Cambridge, MA 02238-2669
617-864-4300
Fax: 617-864-9954
E-mail: deb@alceon.com

Christopher Frey
Assistant Professor
Department of Civil Engineering
North Carolina State University
P.O. Box 7908
Raleigh, NC 27695-7908
919-515-1155
Fax: 919-515-7908
E-mail: frey@eos.ncsu.edu

Susan Griffin
Environmental Scientist
Superfund Remedial Branch
Hazardous Waste
Management Division
U.S. Environmental
Protection Agency
999 18th Street (8EPR-PS)
Suite 500
Denver, CO 80202-2466
303-312-6651
Fax: 303-312-6065
E-mail: griffin.susan@
epamail.epa.gov

Bruce Hope
Environmental Toxicologist
Oregon Department of
Environmental Quality
811 Southwest 6th Avenue
Portland, OR 97204
503-229-6251
Fax: 503-229-6977
E-mail: hope.bruce@deq.state.or.us

William Huber
President
Quantitative Decisions
539 Valley View Road
Merion, PA 19066
610-771-0606
Fax: 610-771-0607
E-mail: whuber@quantdec.com

Robert Lee
Risk Analyst
Golder Associates, Inc.
4104 148th Avenue, NW
Redmond, WA 98052
206-367-2673
Fax: 206-616-4875
E-mail: rclee@u.washington.edu

David Miller
Chemist
Office of Pesticide Programs
Health Effects Division
U.S. Environmental
Protection Agency
401 M Street, SW (7509)
Washington, DC 20460
703-305-5352
Fax: 703-305-5147
E-mail: miller.david@
epamail.epa.gov

Samuel Morris
Environmental Scientist
Deputy Division Head
Brookhaven National Laboratory
Building 815
815 Rutherford Avenue
Upton, NY 11973
516-344-2018
Fax: 516-344-7905
E-mail: morris3@bnl.gov

Jacqueline Moya
Environmental Engineer
National Center for
Environmental Assessment
Office of Research and Development
U.S. Environmental
Protection Agency
401 M Street, SW (8623D)
Washington, DC 20460
202-564-3245
Fax: 202-565-0052
E-mail: moya.jacqueline@
epamail.epa.gov

Christopher Portier
Chief
Laboratory of Computational
Biology and Risk Analysis
National Institute of
Environmental Health Sciences
P.O. Box 12233 (MD- A306)
Research Triangle Park, NC 27709
919-541-4999
Fax: 919-541-1479
E-mail: portier@niehs.nih.gov

P. Barry Ryan
Professor
Exposure Assessment and
Environmental Chemistry
Rollins School of Public Health
Emory University
1518 Clifton Road, NE
Atlanta, GA 30322
404-727-3826
Fax: 404-727-8744
E-mail: bryan@sph.emory.edu

Brian Sassaman
Bioenvironmental Engineer
U.S. Air Force
DET1, HSC/OEMH
2402 E Drive
Brooks Air Force Base, TX 78235-
5114
210-536-6122
Fax: 210-536-1130
E-mail: brian.sassaman@
guardian.brooks.af.mil

Ted Simon
Toxicologist
Federal Facilities Branch
Waste Management Division
U.S. Environmental Protection Agency
Atlanta Federal Center
61 Forsyth Street, SW
Atlanta, GA 30303-3415
404-562-8642
Fax: 404-562-8566
E-mail: simon.ted@epamail.epa.gov

Mitchell J. Small
Professor
Departments of Civil &
Environmental Engineering and
Engineering
& Public Policy
Carnegie Mellon University
Porter Hall 119, Frew Street
Pittsburgh, PA 15213-3890
412-268-8782
Fax: 412-268-7813
E-mail: ms35@andrew.cmu.edu

Edward Stanek
Professor of Biostatistics
Department of Biostatistics
and Epidemiology
University of Massachusetts
404 Arnold Hall
Amherst, MA 01003-0430
413-545-4603
Fax: 413-545-1645
E-mail:
stanek@schoolph.umass.edu

Alan Stern
Acting Chief
Bureau of Risk Analysis
Division of Science and Research
New Jersey Department of
Environmental Protection
401 East State Street
P.O. Box 409
Trenton, NJ 08625
609-633-2374
Fax: 609-292-7340
E-mail: astern@dep.state.nj.us

Paul White
Environmental Engineer
National Center for
Environmental Assessment
Office of Research and Development
U.S. Environmental
Protection Agency
401 M Street, SW (8623D)
Washington, DC 20460
202-564-3289
Fax: 202-565-0078
E-mail: white.paul@epamail.epa.gov

(over)

EPA United States
Environmental Protection Agency
Risk Assessment Forum

Workshop on Selecting Input Distributions for Probabilistic Assessment

U.S. Environmental Protection Agency
New York, NY
April 21-22, 1998

Final List of Observers

Samantha Bates
Graduate Student/
Research Assistant
Department of Statistics
University of Washington
Box 354322
Seattle, WA 98195
206-543-8484
Fax: 206-685-7419
E-mail: sam@stat.washington.edu

Steve Chang
Environmental Engineer
Office of Emergency and
Remedial Response
U.S. Environmental
Protection Agency
401 M Street, SW (5204G)
Washington, DC 20460
703-603-9017
Fax: 703-603-9103
E-mail: chang.steve@
epamail.epa.gov

Helen Chernoff
Senior Scientist
TAMS Consultants, Inc.
655 Third Avenue
New York, NY 10017
212-867-1777
Fax: 212-697-6354
E-mail: hchernoff@
tamsconsultants.com

Christine Daily
Health Physicist
Radiation Protection &
Health Effects Branch
Division of Regulatory Applications
U.S. Nuclear Regulatory Commission
(T-9C24)
Washington, DC 20555
301-415-6026
Fax: 301-415-5385
E-mail: cxd@nrc.gov

Emran Dawoud
Human Health Risk Assessor
Toxicology and Risk Analysis Section
Life Science Division
Oak Ridge National Laboratory
1060 Commerce Park Drive (MS-
6480)
Oak Ridge, TN 37830
423-241-4739
Fax: 423-574-0004
E-mail: dawoudea@ornl.gov

Audrey Galizia
Environmental Scientist
Program Support Branch
Emergency and Remedial
Response Division
U.S. Environmental Protection Agency
290 Broadway
New York, NY 10007
212-637-4352
Fax: 212-637-4360
E-mail: galizia.audrey@
epamail.epa.gov

Ed Garvey
TAMS Consultants
300 Broadacres Drive
Bloomfield, NJ 07003
973-338-6680
Fax: 973-338-1052
E-mail: egarvey@
tamsconsultants.com

Gerry Harris
UMDNJ-RWJMS
UMDNJ-EOSHI
Rutgers University
170 Frelinghuysen Road - Room 234
Piscataway, NJ 08855-1179
732-235-5069
E-mail: gharris@gpph.rutgers.edu

David Hohreiter
Senior Scientist
BBL
6723 Towpath Road
P.O. Box 66
Syracuse, NY 13214
315-446-9120
Fax: 315-446-7485
E-mail: dh%bbl@mcimail.com

Printed on Recycled Paper

(over)

Nancy Jafolla
Environmental Scientist
Hazardous Waste
Management Division
U.S. Environmental
Protection Agency
841 Chestnut Building (3H541)
Philadelphia, PA 19107
215-566-3324
E-mail: jafolla.nancy@
epamail.epa.gov

Alan Kao
Senior Science Advisor
ENVIRON Corporation
4350 North Fairfax Drive - Suite 300
Arlington, VA 22203
703-516-2308
Fax: 703-516-2393

Steve Knott
Executive Director, Risk
Assessment Forum
Office of Research and Development
National Center for
Environmental Assessment
U.S. Environmental
Protection Agency
401 M Street, SW (8601-D)
Washington, DC 20460
202-564-3359
Fax: 202-565-0062
E-mail:
knott.steve@epamail.epa.gov

Stephen Kroner
Environmental Scientist
Office of Solid Waste
U.S. Environmental
Protection Agency
401 M Street, SW (5307W)
Washington, DC 20460
703-308-0468
E-mail: kroner.stephen@
epamail.epa.gov

Anne LeHuray
Regional Risk Assessment Lead
Foster-Wheeler
Environmental Corporation
8100 Professional Place - Suite 308
Lanham, MD 20785
301-429-2116
Fax: 301-429-2111
E-mail: alehuray@fwenc.com

Toby Levin
Attorney
Advertising Practices
Federal Trade Commission
601 Pennsylvania Avenue, NW
Suite 4110
Washington, DC 20852
202-326-3156
Fax: 202-326-3259

Lawrence Myers
Statistician
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709
919-541-6932
Fax: 919-541-5966
E-mail: lem@rti.org

Marian Olsen
Environmental Scientist
Technical Support Section
Program Support Branch
Emergency and Remedial
Response Division
U.S. Environmental Protection Agency
290 Broadway
New York, NY 10007
212-637-4313
Fax: 212-637-4360
E-mail:
olsen.marion@epamail.epa.gov

Lenwood Owens
President
Boiler Servicing
1 Laguardia Road
Chester, NY 10918

Zubair Saleem
Office of Solid Waste
U.S. Environmental Protection Agency
401 M Street, SW (5307W)
Washington, DC 20460
703-308-0467
Fax: 703-308-0511
E-mail: saleem.zubair@
epamail.epa.gov

Swati Tappin
Research Scientist
New Jersey Department of
Environmental Protection
401 East State Street
P.O. Box 413
Trenton, NJ 08625
609-633-1348

Joan Tell
Senior Environmental Scientist
Exxon Biomedical
Mettlers Road (CN 2350)
East Millston, NJ 08875
732-873-6304
Fax: 732-873-6009
E-mail: joan.tell@exxon.sprint.com

Bill Wood
Executive Director, Risk
Assessment Forum
Office of Research and Development
National Center for
Environmental Assessment
U.S. Environmental
Protection Agency
401 M Street, SW (8601-D)
Washington, DC 20460
202-564-3358
Fax: 202-565-0062
E-mail: wood.bill@epamail.epa.gov

APPENDIX C

AGENDA

EPA United States
Environmental Protection Agency
Risk Assessment Forum

Workshop on Selecting Input Distributions for Probabilistic Assessment

U.S. Environmental Protection Agency
New York, NY
April 21-22, 1998

Agenda

Workshop Chair: **Christopher Frey**
North Carolina State University

TUESDAY, APRIL 21, 1998

8:00AM **Registration/Check-In**

9:00AM **Welcome Remarks**
Representative from Region 2, U.S. Environmental Protection Agency (U.S. EPA), New York, NY

9:10AM **Overview and Background**
Steve Knott, U.S. EPA, Office of Research and Development (ORD), Risk Assessment Forum, Washington, DC

9:30AM **Workshop Structure and Objectives**
Christopher Frey, Workshop Chair

9:45AM **Introduction of Invited Experts**

10:00AM **Presentation: Issue Paper #1 - Evaluating Representativeness of Exposure Factors Data**
Jacqueline Moya, U.S. EPA, National Center for Environmental Assessment (NCEA), Washington, DC

10:15AM **Presentation: Issue Paper #2 - Empirical Distribution Functions and Non-Parametric Simulation**
Tim Barry, U.S. EPA, NCEA, Washington, DC

10:30AM BREAK

10:45AM **Charge to the Panel**
Christopher Frey, Workshop Chair

11:00AM **Discussion on Issue #1: Representativeness**

12:00PM LUNCH

(over)

TUESDAY, APRIL 21, 1998 (continued)

1:30PM **Discussion on Issue #1 Continues**

3:00PM BREAK

3:15PM **Discussion on Issue #1 Continues**
Christopher Frey, Workshop Chair

4:15PM **Observer Comments**

4:45PM **Review of Charge for Day Two**
Christopher Frey, Workshop Chair

- Writing Assignments

5:00PM ADJOURN

WEDNESDAY, APRIL 22, 1998

8:30AM **Planning and Logistics**
Christopher Frey, Workshop Chair

8:40AM **Summary of Discussion on Issue #1**

10:00AM BREAK

10:15AM **Discussion on Issue #2: Empirical Distribution Functions and Resembling Versus Parametric Distributions**

12:00PM LUNCH

1:30PM **Discussion on Issue #2 Continues**

3:00PM BREAK

3:15PM **Summary of Discussion on Issue #2**
Christopher Frey, Workshop Chair

- Writing Assignments/Session

4:15PM **Observer Comments**

4:45PM **Closing Remarks**

5:00PM ADJOURN

APPENDIX D

WORKSHOP CHARGE

Workshop on Selecting Input Distributions for Probabilistic Assessment

U.S. Environmental Protection Agency
New York, NY
April 21-22, 1998

Charge to Experts/Discussion Issues

This workshop is being held to discuss issues associated with the selection of probability distributions to represent exposure factors in a probabilistic risk assessment. The workshop discussions will focus on generic technical issues applicable to any exposure data. It is not the intent of this workshop to formulate decisions specific to any particular exposure factors. Rather, the goal of the workshop is to capture a discussion of generic issues that will be informative to Agency assessors working with a variety of exposure data.

On May 15, 1997, the U.S. Environmental Protection Agency (EPA) Deputy Administrator signed the Agency's "Policy for Use of Probabilistic Analysis in Risk Assessment." This policy establishes the Agency's position that "such probabilistic analysis techniques as Monte Carlo Analysis, given adequate supporting data and credible assumptions, can be viable statistical tools for analyzing variability and uncertainty in risk assessments." The policy also identifies several implementation activities designed to assist Agency assessors with their review and preparation of probabilistic assessments. These activities include a commitment by the EPA Risk Assessment Forum (RAF) to organize workshops or colloquia to facilitate the development of distributions for exposure factors.

In the summer of 1997, a technical panel, convened under the auspices of the RAF, began work on a framework for selecting input distributions for use in Monte Carlo analyses. The framework emphasized parametric methods and was organized around three fundamental activities: selecting candidate theoretical distributions, estimating the parameters of the candidate distributions, and evaluating the quality of the fit of the candidate distributions. In September of 1997, input on the framework was sought from a 12 member panel of experts from outside of the EPA. The recommendations of this panel include:

- expanding the framework's discussion of exploratory data analysis and graphical methods for assessing the quality of fit,
- discussing distinctions between variability and uncertainty and their implications,
- discussing empirical distributions and bootstrapping,
- discussing correlation and its implications,
- making the framework available to the risk assessment community as soon as possible.

Subsequent to receiving this input, some changes were made to the framework and it was applied to selecting distributions for three exposure factors: water intake per body weight, inhalation rate, and residence time. The results of this work are presented in the attached report entitled "Development of Statistical Distributions for Exposure Factors."

Applying the framework to the three exposure factors highlighted several issues. These issues resolved into two broad categories: issues associated with the representativeness of the data, and issues associated with using the empirical distribution function (or resampling techniques) versus using a theoretical parametric distribution function. Summaries for these issues are presented in the attached issue papers. These issues will be the focal point for discussions during this workshop. The following questions are intended to help structure and guide these discussions. In addressing these questions, workshop participants are asked to consider: what do we know today that can be applied to answering the question or providing additional guidance on the topic; what short term studies (e.g., numerical experiments) could be conducted to answer the question or provide additional guidance; and what longer term research may be needed to answer the question or provide additional guidance.

Representativeness (Issues Paper #1)

1) The Issue Paper

Checklists I through IV in the issue paper present a framework for characterizing and evaluating the representativeness of exposure data. This framework is organized into three broad sets of questions: questions related to differences in populations, questions related to differences in spatial coverage and scale, and questions related to differences in temporal scale. Do these issues cover the most important considerations for representativeness? Are the lists of questions associated with each issue complete? If not, what questions should be added?

In a tiered approach to risk assessment (e.g., a progression from simpler screening level assessments to more complex assessments), how might the framework be tailored to each tier? For example, is there a subset of questions that adequately addresses our concerns about representativeness for a screening level risk assessment?

2) Sensitivity

The framework asks how important are (or how sensitive is the analysis to) population, spatial, and temporal differences between the sample (for which you have the data) and the population of interest. For example, to what extent do these differences affect our estimates of the mean and variance of the population and what is the magnitude and direction of these effects?

What guidance can be provided to help answer these questions? What sources of information exist to help with these questions? Having answered these questions what are the implications for the use of the data (e.g., use of the data may be restricted to screening level assessments in certain circumstances)? What differences could be considered critical (i.e., what differences could lead to the conclusion that the assessment can't be done without the collection of additional information)?

3) Adjustments

The framework asks, is there a reasonable way of adjusting or extrapolating from the sample (for which you have data) to the population of interest in terms of the population, spatial, and temporal characteristics? If so, what methods should be used? Is there adequate information available to implement these methods?

What guidance can be provided to help answer these questions? Can exemplary methods for making adjustments be proposed? What sources of information exist to help with these questions? What research could address some of these issues?

Section 5 of the issue paper on representativeness describes methods for adjustments to account for differences in population and temporal scales. What other methods exist? What methods are available for spatial scales? Are there short-term studies that can be done to develop these methods further? Are there data available to develop these methods further? Are there numerical experiments (e.g., simulations) that can be done to explore these methods further?

Empirical Distribution Functions and Resampling Versus Parametric Distributions (Issues Paper #2)

1) Selecting the EDF or PDF

What are the primary considerations for assessors in choosing between the use of theoretical parametric distribution functions (PDFs) and empirical distribution functions (EDFs) to represent an exposure factor? Do the advantages of one method significantly outweigh the advantages of the other? Is the choice inherently one of preference? Are there situations in which one method is clearly preferred over the other? Are there circumstances in which either method of representation should not be used?

2) Goodness of Fit

On what basis should it be decided whether or not a data set is adequately represented by a fitted analytic distribution? What role should the goodness-of-fit test statistic play (e.g., chi-square, Kolmogorov-Smirnov, Anderson-Darling, Cramer-von Mises, etc.)? How should the level of significance, i.e., p-value, of the goodness of fit statistic be chosen? What are the implications or consequences for exposure assessors when acceptance/rejection is dependent on the goodness of fit statistic chosen and an arbitrary level of statistical significance? What role should graphical examination of the quality of fit play in the decision as to whether a fit is acceptable or not?

When the only data readily available are summary statistics (e.g., selected percentiles, mean, and variance), are fits to analytic distributions based on those summary statistics acceptable? Should any limitations or restrictions be placed in these situations?

When the better known theoretical distributions (e.g., lognormal, gamma, Weibull, log-logistic, etc.) cannot provide an acceptable fit to a particular set of data, is there value in testing the fit of the more flexible generalized distributions (e.g., the generalized gamma and generalized F distributions) even though they are considerably more complicated and difficult to work with?

3) Uncertainty

Are there preferred methods for assessing uncertainty in the fitted parameters (e.g., methods based on maximum likelihood and asymptotic normality, bootstrapping, etc.)?

APPENDIX E

BREAKOUT SESSION NOTES

APPENDIX E

SMALL GROUP DISCUSSIONS/BRAINWRITING SESSIONS

During the workshop, the experts worked at times in smaller groups to discuss specific technical questions. Some of these sessions involved open discussions. Other sessions involved "brainwriting," during which individuals captured their thoughts on paper, in sequence, and then discussed similar and/or opposing views within each group. The outcomes of these sessions were captured by group rapporteurs and individual group members and are summarized below. This summary represents a transcription of handwritten notes and are, as such, considered rough working notes. Information from these smaller group discussions was presented and deliberated in the plenary session, and partially forms the basis of the points presented in the main text of this report.

What information is required to fully specify a problem definition?

- Population at risk
- Sample under study (include biases)
- Spatial extent of exposure—micro, meso, macro scale
- Exposure-dose relationship
- Dose-response–risk relationship
- Temporal extent (hours, days, months, years)
- Temporal variability about trend
- What is the "acceptable error"?
  - yes/no
  - categorization
  - continuous
  - quantitative
- Variability/uncertainty partitioning
  - not needed
  - desirable
  - mandatory
- User of output
  - scientific community
  - regulatory community
  - general public

One expert noted that the "previous problem definition" forces the blurring of the boundaries between modeling and problem description—for example, many may not consider the dose-exposure–risk relationship to be part of the problem definition.

Another expert asked, "How much information do we have to translate from measured value to population of concern?" He described the population of concern, surrogate population, individuals sampled from the surrogate population, and how well measured value represents true value. Another agreed, emphasizing the importance of temporal, spatial, and temporal-spatial representativeness (e.g., Idaho potatoes versus Maine potatoes).

Other issues in problem definition include:

- In the context of environmental remediation, a problem is defined in terms of what level of residual risk can be left on the site. The degree of representativeness needed is dependent on the land use scenario.

Several alternative scenarios of future land use, population, etc. might be defined and analyzed. Problem definition might include establishing budget limits (for assessment and remediation); this might dictate limits on future land use and the need for evaluation.

- A problem needs to be specified in space (location), time (over what duration), and whom (person or unit). Some of these definitions may be concrete (e.g., in terms of spatial locations around a site) while some may be more vague, such as persons who live on a brownfield site (which may change over time with mobility, new land use, etc.). The problem addresses a future context, and must therefore be linked to observable data by a model/set of assumptions. The problem definition should include these models (no population change over time) or assumptions (exposure calculated over 50- year duration/time frame).

- One must define the health outcome being targeted (e.g., acute vs. cancer vs. developmental).

Define how you will link the exposure measure to a model for hazard and/or risk (margin of exposure has different data needs from an estimate of population risk). Also, one should consider the *type* of observation being evaluated (blood measurements vs. dietary vs. ecological). This is more likely to have an impact on the representativeness of the data sample than anything else.

Define the target risk level; this will dictate what kind of data will be necessary.

Another panelist agreed these are important points but questioned, however, whether these factors were part of problem definition.

- Specify the scope and purpose of the assessment (e.g., regulatory decision, set cleanup standards, etc.)

- Determining how much error we are willing to live with will determine how representative the data are.

- Specify the population of concern (who they are, where they live, what kinds of activities they are involved with).

- Problem definition is the most critical part of the process, and all stakeholders should be involved as much as possible. If the stakeholders come to a common understanding of the objectives of the process, the situation becomes focused.

- Although EPA has provided much guidance for problem definition (DQOs, DQAs, etc.), what data are necessary (and to what extent it must be representative) is a function of each individual problem. Certain basic questions are common to all problem definitions (who, what, when, how); the degree to which each basic question is important is a function of the actual problem/situation.

Decision performance requirements: What is acceptable at a specific site for a specific problem (i.e., what is the degree of decision error)? An answer to this question should be decided up front as much as possible to alleviate "bias" concerns.

- Attributes of the exposed population are key issues:

  - Who are they?
  - What are their activities/behaviors?
  - Where are they?
  - When do they engage in activities and for how long?
  - Why are certain activities performed?

- The potential imprecision of "national" populations seems significant. Scale is important; maybe regional is as large as it gets.

- If representativeness is a property of the population, then we should focus on methods for collecting more specific data.

- Variability within a super-population (e.g., a national study) provides useful, quantifiable bounds to potential bias and gives an upper bound on the variability that could be found in a subpopulation. This suggests that there are quantitative ways to guide the use "reduce sparingly."

- The assessor needs to ask the following questions: Is a risk assessment necessary? What is the level of detail needed for the decision at hand? What is the scope of the problem? For example,

  - Who is at risk?
  - Who has standing [e.g., stakeholders]?
  - Who has special concerns?
  - What is of concern?
  - When are people exposed? (timeframe [frequency and duration], chronic vs. acute, level of time steps needed)
  - Where are people exposed—spatial considerations; scope of the problem (national, regional, site?)
  - How are people exposed?

- The time step used in the model must be specified. The assessor must distinguish between distribution needed for a one-day time step as compared to a one-year time step. Some models may run at different time steps (e.g., drinking water at a one-week time step to include seasonal variation; body weight at a one-year time step to include growth of a child.)

- Consideration of a tiered approach is important in problem formulation. How are data to be used? If data are to be used in a screening manner, then conservativeness is even more important than representativeness. If more than a screening assessment is proposed, the assessor should consider *what is the value added* from more complex analyses (site-specific data collection, modeling, etc.).

- As probabilistic methods continue to be developed, it will become increasingly important to specify constraints in distribution. Boundaries exist. For example, no person can eat multiple food groups at the 95th percentile.

- Two panelists noted that tiered approaches would not change the problem definition. Generally, the problem is: Under an agreed set of exposure conditions, will the population of concern experience unacceptable risks? This question would not change with a more or less sophisticated (tiered) assessment.

- When evaluating unknown future population characteristics, we are dealing with essentially unknown conditions. It is not feasible, therefore, to have as a criterion that additional information will not significantly change the outcome of the analysis. Instead, the problem needs to be defined in terms of a precise definition of population (in time and space) which is to be protected. To the extent that this is uncertain, it needs to be defined in a generalized, generic manner.

- Considerations of the "external" representativeness of the data to the population of concern is absolutely critical for "on the ground" risk assessments. The "internal" validity of the data is often a statistical question. It seems more important to ensure that the outcome of the assessment will not change based on the consideration of "external" representativeness of the data set to the population of concern.

What constitutes (lack of) representativeness?

General
The issue of data representativeness begs the question "representative of what?" In many (most?) cases, we are working backwards, using data in hand for purposes that may or may not be directly related to the reason the data were collected in the first place. Ideally, we would have a well-posed assessment problem with well-defined assessment endpoints. From that starting point, we would collect the relevant data necessary for good statistical characterization of the key exposure factors.

More generally, we are faced with the question, "Can I use these data in my analysis?" To make that judgment fairly, we would have to go through a series of questions related to the data itself and to the use we intend to make of the data. We usually ignore many of these questions, either explicitly or implicitly. The following is an attempt at listing the issues that ought to affect our judgment of data relevance.

Sources of Variability and Uncertainty Related to the Assessment of Data Representativeness
EPA policy sets the standard that risk assessors should seek to characterize central tendency and plausible upper bounds on both individual risk and population risk for the overall target population as well as for sensitive subpopulations. To this extent, data representativeness cannot be separated from the assessment endpoint(s). The following outlines some of the key elements affecting data representativeness. The elements are not mutually exclusive.

Exposed Population

general target population
particular ethnic group
known sensitive subgroup (children, elderly, asthmatics, etc.)
occupational group (applicators, etc.)
age group (infant, child, teen, adult, whole life)
sex
activity group (sport fishermen, subsistence fishermen, etc.)

Geographic Scale, Location

trends (stationary, non-stationary behaviors)
past, present, future exposures
lifetime exposures
less-than-lifetime exposures (hourly, daily, weekly, annually, etc.)
temporal characteristics of source(s), continuous, intermittent, periodic, concentrated (spike), random

Exposure Route

inhalation
ingestion (direct, indirect)
dermal (direct) contact (by activity, e.g., swimming)
multiple pathways

Exposure/Risk Assessment Endpoint

cancer risk
non-cancer risk (margin of exposure, hazard index)
potential dose, applied dose, internal dose, biologically effective dose
risk statistic
mean, uncertainty percentile of mean
percentile of a distribution (e.g., 95th percentile risk)
uncertainty percentile of variability percentile (upper credibility limit on 95th percentile risk)
plausible worst case, uncertainty percentile of plausible worst case

Data Quality Issues

direct measurement, indirect measurement (surrogates)
modeling uncertainties
measurement error (accuracy, precision, bias)
sampling error (sample size, non-randomness, independence)
monitoring issues (short-term, long-term, stationary, mobile)

■ Almost all data used in risk assessment is not representative in one or more ways. What is important is the effect the lack of representativeness has on the risk assessment in question. If the water pathway, for example, is of minor concern, it will not matter if the water-consumption rate distribution is not representative.

A lack of representativeness could mean the risk assessment results fail to be protective of public health or grossly overestimate risks.

The Issue Paper is helpful in describing the ways in which distributions can be nonrepresentative. It can guide the selection of the input distributions.

- Representativeness needs to be considered in the context of the decision performance requirements. Factors that could have a major impact in terms of one problem/site need not have the same impact across all problems/sites. Decision performance requirements should therefore be considered with problem-site-specific goals and objectives factored into the process.

- The definition of representativeness depends on how much error we are willing to live with. What is "good enough" will be case specific. Going through some case studies using assessments done for different purposes can shed some light on defining representativeness. "With regard to exposure factors, we [EPA] need to do a better job at specifying or providing better guidance on how to use the data that are available." For example, the soil ingestion data for children are limited, but may be good enough to provide an estimate of a mean. The data are not good enough to support a distribution or a good estimate of a high-end value.

- Representativeness measures the degree to which a sample of values for a given endpoint accurately and precisely (*adequately*) describes the value(s) of that endpoint likely to be seen in a target population.

- A number of issues relate to the lack of representativeness which one can use to decide upon use of a sample in a given case: The context of the observation is important. In addition to those mentioned in the Issues Paper (demographic, technical, social), other concerns include what is being measured: environmental sample (water, air, soil) versus human recall (diet) versus tissue samples in humans (e.g., blood). In most cases, provided good demographic and social information is available on key issues associated with the exposure, adjustment can be made to make a sample representative for a new population. Technical issues sometimes must be "guessed" from one sample to another (key issues like different or poor analytic techniques, altered consumption rates, etc.).

- A sample should not be used if it is flawed due to one of the following factors:

  1) inappropriate methods (sample design and technical methods)
  2) lack of descriptors (demographic, technical, social) to make adjustments
  3) inadequate size for target measure

  The above applies to the internal analysis of a sample. Human recall includes behavioral activities (e.g., time spent outdoors or indoors, number of days away from site).

- Identifying differences (as defined by the final objective) between characteristics of the subject population and the surrogate population will generally be subjective because there is usually no data for the subject population. Differences might be due to socioeconomic differences, race, or climate. Lack of representativeness should not be "too rigid" partly due to uncertainties and partly because the subject population usually includes a future population that is even less well defined than the current population.

The surrogate population may overlap (as in age/sex distribution) with the target population. A context is needed to determine what constitutes "lack of representativeness." For example, if soil ingestion is not related to gender, then while the surrogate population may be all female, it may not imply that the estimates from the surrogate population cannot be used for a target population (including males and females). Bottom line: the factor being represented (such as gender) needs to be related to the outcome (soil ingestion) before the non-representativeness is important. Lack of representativeness "depends" in this sense on the association.

Another panelist expanded on the above, noting that the outcome determines the representativeness of the surrogate data set. If in the eyes of the "beholder" the data are "equivalent" they represent the actual population well. Defining representativeness is like defining art. One cannot describe it well; it is easily recognized but recognition is observer-dependent. We should strive to remove subjectivity as best as possible without making inflexible choices.

- Representativeness suggests that our exposure/risk model results are a reasonable approximation of reality. At minimum, they pass a straight-face test. Representativeness could therefore be assessed via model calibrations and validation.

- Representativeness often cannot be addressed unless an expert-judgment-based approach is used. It requires brainstorming based upon some knowledge of how the target population may differ from the surrogate one. In the long run, collection of more data is needed to reduce the non-representativeness of those distributions upon which decisions are based.

- Define the characteristics to be examined, define the population to be evaluated, select a statistically significant sample that reflects defined characteristics of the population (another expert noted that statistical significance has little relevance to the problem of representativeness—the issue is the degree of uncertainty or bias). Ensure randomness of a sample to capture the entire range of population characteristics. (Another noted that the problem is that we usually don't have such a sample but have to make a decision or take action now. If we can quantitatively evaluate representativeness, then we can at least make objective determinations of whether this lack of representativeness will materially affect the decisions.)

- The degree of bias that exists between a data set or sample and the problem at hand—is the sample even relevant to the problem? Types:

Scenario:	Is a "future residential" scenario appropriate to the problem at hand?
Model:	Is a multiplicative, independent-variable model appropriate?
Variables:	Is a particular study appropriate to the problem? Is it biased? Uncertain?

- Two experts agreed that statistical significance has little relevance to the problem of representativeness. A well-designed controlled randomized study yielding two results can be "representative" of the mean and dispersion, albeit highly imprecise.

- Representativeness exists when the data sample is drawn at random from the population (including temporal and spatial characteristics) of concern, or is a census in the absence of measurement error. This condition is potentially lacking when using surrogate data that are for a population that differs in any way from the population of concern. Important differences include:

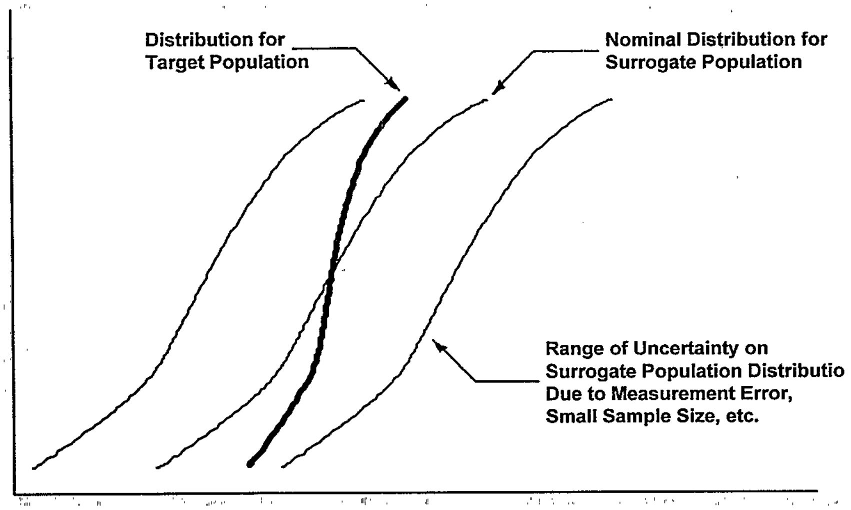
  — characteristics of individuals (e.g., age, sex, etc.)
  — geographic locations
  — averaging time
  — dynamics of population characteristics over the time frame needed in the study
  — large measurement errors

  Non-representativeness poses a problem if we have biases in any statistical interest (i.e., lack of representativeness can lead to biases in the mean, standard deviation, 95th percentile, etc).

  Bias, or lack of accuracy, is typically more important than lack of precision. For example, we can expect some imprecision in our estimate of the 95th percentile of a population characteristic (e.g., intake rate) due to lack of relevant "census" data, but we hope that on average our assessment methods do not produce a bias or systematic error.

  Conversely, if we have a large amount of uncertainty in our estimates for a sample distribution, then it is harder to claim non-representativeness than when a particular distribution for a surrogate is estimated.

  In the following example, the distribution for the surrogate population is non-representative of the target population since it has too wide a variance. However, the uncertainty in the surrogate encompasses outcomes which could include the target population. Thus, in this case it may be difficult to conclude, based upon the wide range of uncertainty, that the surrogate is non-representative.



- Representativeness in a given exposure variable is determined by how well a given data set reflects the characteristics of the population of concern. Known characteristics of the data that distinguish the data set from the population of concern may indicate a need for adjustment. Areas of ignorance regarding the data set and the population of concern should be considered uncertainties. Representativeness or lack thereof should be determined in a brainstorming session among stakeholders. Toxicologists, statisticians, engineers, and others may all have information that bears on the representativeness of the data. Known or suspected difference between the data set and the population of concern diminish representativeness.

- The question as to what constitutes representativeness is contingent on the problem definition—that is, who is to be represented, at what point in time, etc. If the goal is to represent a well-characterized population in the present, representativeness for a given parameter (e.g., drinking water consumption) should be evaluated based on the match of the surrogate data to the data for the population of concern relative to key correlates of the parameter (e.g., for drinking water volume, age, average ambient temperature, etc.). If, on the other hand, the population of concern is not well characterized in the present, or if the intent of the risk assessment is to address risk into the indefinite future, representativeness does not appear to have a clear meaning. The goal in such cases should be to define reasonable screening characteristics of a population at an indefinite point in time (e.g., maximum value, minimum value, estimated 10th percentile, estimated 90th percentile) and select such values from a semi-quantitative analysis of the available surrogate data.

- A representative surrogate sample is one that adds information to the assessment beyond the current state of knowledge. However, both the degree to which it adds information and the remaining uncertainty in the risk characterization must be identified.

- Suggestion: Replace the word representative with "useful and informative."

- A data set is representative of a characteristic of the population if it can be shown that differences between the data set and the population of concern will not change the outcome of the assessment. In practice, a data set should be considered in terms of its similarity and difference to the population of concern and expectations as to how the differences might change the outcome. Of course, these expectations may lead to adjustments in the data set which would make it potentially more representative of the population.

- In part, what degree of comfort the risk assessor/reviewer needs to have for the population under consideration determines how representative data have to be. Also of concern is where in the population of concern observations will take place. Are we comparing data mean or tails (outliers)? What degree of uncertainty and variability between the population of concern and the surrogate data is the assessor willing to live with?

- We may be using the term "representativeness" too broadly. Many of the issues seem to address the "validity" of the study being evaluated. However, keeping with the broad definition, the following apply to internal representativeness:

- *Measurement reliability*. Measurement reliability refers whether the study correctly measures what it set out to measure and provides some basis for evaluating the error in measurement.

- *Bias in sampling.* Bias in sampling presupposes that there is a "population" that was sampled and not just a haphazard collection of observations and measurements.

- *Statistical sampling error.*

The following issues apply to external representativeness:

- Did the study measure what we need to know (e.g., short-term vs. long-term studies). If there is a statistical procedure for translating measurements into an estimate of the needed values, the validity and errors involved must be considered.

- "Representativeness" implies that the sample data is appropriate to another population in an assessment.

What considerations should be included in, added to, or excluded from the checklists?

- Expand to include other populations of concern (e.g., ecological, produce). The issue paper and checklist seem to presuppose that the population of concern is the human population.

- Include more discussion on criteria for determining if question is adequately and appropriately answered.

- Clarify definitions (e.g., internal versus external)

- Include "worked" examples:

  - Superfund-type risk assessment
  - Source-exposure-dose-effect-risk example
  - Include effect of bias, misclassification, and other problems

- Ask if factors are known or suspected of being associated with the outcome measured? Was the distribution of factors known or suspected to be associated with the outcome spanned by the sample data? Focus on outcome of risk assessments (if populations are different, does it make any real difference in the outcome of the assessment?).

- How will the exposures be used in risk assessment? For example, is the sample representative enough to bound the risk?

- In judging the quality of a sample, especially with questionnaire-based data, determine whether a consistency check was put in the forms and the degree to which individual samples are consistent. Risk assessors must be able to review the survey instrument.

- Internal and external lists may each need some reorganization (for example, measurement issues vs. statistical bias and sampling issues for "internal;" extrapolation to a different population vs. reanalysis/reinterpretation of measurement data for "external").

- Is a good set of subject descriptors (covariates such as age, ethnicity, income, education, or other factors that can affect behavior or response) available for both the population sampled and population of concern to allow for correlations and adjustments based on these?

- How valuable would some new or additional data collection be for the population of concern to confirm the degree of representativeness of the surrogate population and better identify and estimate the adjustment procedure?

- What is the endpoint of concern and what decision will be based on the information that is gathered? Since risk assessment involves a tiered approach, checklist should focus around the following type of question: Do I have enough information about population (type, space, time) that allow answering the questions at this tier and is my information complete enough that I can make a management decision? Do I need to go through all of the checklists before I can stop? (Questioning application/implementation)

- The checklists should address how much is known about the population of concern relative to the adaptation of the surrogate data. If the population of concern is inadequately characterized, then the ability to consider the representativeness of the surrogate data is limited, and meaningless adjustment will result.

- One consideration that is missing from the checklists is the fact that risk assessments are done for a variety of purposes. A screening level assessment may not need the level of detail that the checklists include. The checklists should be kept as simple and short as possible, trying to avoid redundancy.

- The checklist should be flexible enough to cover a variety of different problems and should be only a guide on how to approach the problem. The more considerations included the better.

- Guidance is needed on how to address overlap of the checklists. For example, when overlap exists (e.g., in some spatial and temporal characteristics), which questions in the checklist are critical? The guidance could use real life case studies to help focus the risk assessor on the issues that are critical to representativeness.

- Move from a linear checklist format to a flowchart/framework centered around the "critical" elements of representativeness.

- Fold in nature of tiered analysis. The requirements of a screening level assessment must be different from those of a full-blown risk assessment.

- Identify threshold (make or break) issues to the extent possible (i.e., minimum requirements).

- When biases due to lack of representativeness are suspected, how can we judge which direction those biases take (high or low?).

- Include a "box" describing cases when "nonrepresentative" and "inadequate" will need to be used in a risk assessment (which is common)....Figure 1?

- Define ambiguous terms, such as "reasonable" and "important."

- Make checklist more than binary (yes, no)—allow for qualitative evaluation of data.

  Key questions: Can data be used at all? If so, do we have a great deal of confidence in it or not? Is data biased high or low? Can data be used in a quantitative, semi-quantitative, or only a qualitative manner? Standards according to which checklist items are evaluated should be consistent with stated objective (e.g., a screening assessment will require less stringent evaluation of data set than a site assessment where community concerns or economic costs are critical issues).

- Allow for professional judgement and expert elicitation.

- What are the representativeness decision criteria? Data only have to be good enough for the problem at hand; there are no perfect data. List some considerations pertaining to the acceptance/rejection criteria.

- The 95th percentile of each input distribution is not needed to forecast risk at the 95th percentile with high accuracy and low uncertainty.

- What is the study population doing? (i.e., were the sample population and study population engaged in similar activities?) Consider how their behavior affects ability to represent.

- Combine Checklists II, III, and IV into one.

- Distinguish between marginal distributions vs. joint distributions vs. functional relationships.

- Distinguish variability from uncertainty. Add a crisp definition of each (e.g., Burmaster's premeeting comments).

- Add explicit encouragement and positive incentives to collect and analyze new data.

- Add an explicit statement that the agency encourages the development and use of new methods and that nothing in this guidance should be interpreted as blocking the use of alternative or new methods.

- Add an explicit statement that it is always appropriate to combine information from several studies to develop a distribution for an exposure factor. (This also applies to toxicology and the development of distributions for reference doses and cancer slope factors.)

How can one perform a sensitivity analysis to evaluate the implications of non-representativeness? How do we assess the importance of non-representativeness?

- The assessor should ask, "under a range of plausible adjustments from the surrogate population to the population of concern, does (or can) the risk management decision change?" That is, do these particular assumptions and their uncertainty matter? (among all others)

  Representativeness is often not that important, because risk management decisions are usually not designed to protect just the current population at a particular location, but a range of possible target populations (e.g., future site or product users) under different possible scenarios.

- Theoretically, we can come up with a "perfect" risk assessment in terms of representativeness, but if the factor(s) being evaluated is not important, then the utility of this perfectly representative data is limited. The important question to ask is: If one is wrong, what are the consequences, and what difference do the decision errors make in the estimate of the parameter being evaluated?

  The question of data representativeness can be asked absent the context/model/parameter or it can be asked in the context of a decision or analysis (are the data adequate?).

  The key is placing bounds on the use of the data. Assessments should be put in context and the level at which surrogate data may be representative. It should be defined in the context of the purpose of the original study. Two other factors are critical: sensitivity and cost/resource allocation. The question, therefore, is situation-specific.

- A sensitivity analysis can be conducted in the context of the following tiered approach. The importance of a parameter (as evidenced by a sensitivity analysis) is determined first, making the representativeness or non-representativeness of the non-sensitive parameters unimportant.

- Representativeness is not a standard statistical term. Statistical terms that may be preferable include bias and consistency.

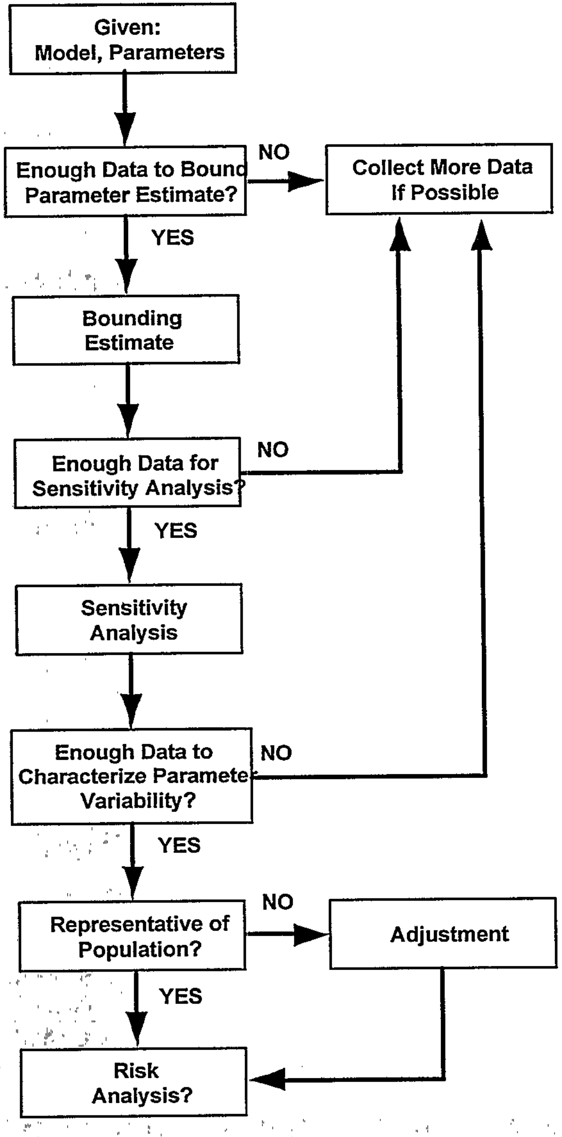
  When evaluating the importance of non-representativeness, one needs to evaluate the uncertainty on the data set and on the individual. At the first level the assessor may choose a value biased high (could be a point value or a distribution that is shifted up). At the second level, can use an average, but must still be sensitive to whether acute or chronic effects are being evaluated. When looking at the individual sample it is more important to have a representative sample because the relevant data are in the tails (more important for acute toxicity). When using a mixture, representativeness is less of a problem.

Adjustments

- Take more human tissue samples to back calculate—this makes local population happier. Determine the need for cleanup based on tissue sample findings.

- Re-do large samples (e.g., food consumptions, tapwater consumption).

- Look at demographics, etc. and determine the most sensitive factor(s).


Given:
Model, Parameters
Enough Data to Bound Parameter Estimate?
NO
Collect More Data If Possible
YES
Bounding Estimate
Enough Data for Sensitivity Analysis?
NO
YES
Sensitivity Analysis
Enough Data to Characterize Parameter Variability?
NO
YES
Representative of Population?
NO
Adjustment
YES
Risk Analysis?

- Use a general model. Discuss with stakeholders the degree of inclusion in general. Adjust the model with survey data if it is not applicable to stakeholder. Use a special model for subpopulations if necessary.

- "Change of support" analysis; time-series analysis — non-CERCLA, important to the Food Quality Protection Act/

- Conduct three-day surveys with year-long adjustments.

- Hypothesis methods will work, but need to be tested.

- The group recommended holding a workshop for experts in related fields to share *existing* theory and methods on adjustment (across fields).

- General guidelines for adjustments will be acceptable, but often site-specific needs dictate what adjustments must be made.

- Example adjustment:

  Fish consumption: If you collect data 3 days per week, you may miss those who might eat less—a case of inter- versus intra-individual variability.

- Adjustment is often difficult because of site specifics and evaluator bias or professional judgement.

- Sometimes it is not possible to adjust. Using an alternate surrogate data set makes it possible to set some plausible bounds to perform a screening risk assessment.

- Stratify data to see if any correlation exists.

- Start with brainstorming.

- Regression relationship versus threshold.

- Covariance; good statistical power to sample population.

- Correlation is equivalent to regression analysis as long as you keep the residual (Bayesian presentation).

- Instead of looking at the population, look at the individual (e.g., breathing rates or body weight for individuals from ages 0 to 30) to establish correlations.

- What if the population was misrepresented? For example, population of concern is sport fishermen but the national data represent other types of fishermen.

Set up a hierarchy:

- do nothing (may fall out when bounded)
- conservative/plausible upper bound
- use simple model to adjust the data (may be worth the effort if credibility issues are dealt with)
- resample/collect more data

Before considering a bounding approach (model development), consider if refining is necessary or cost/beneficial.

Are there situations in which "g-estimates" are worthwhile?

- What is gained by making adjustments?

  Short-term studies overestimate variability because they do not account for interindividual variability (upper tail is overstated).

- Can we estimate the direction of biases when populations are mismatched?

  If the bias is conservative, then we are being protective. But what if the bias is nonconservative (e.g., drinking water in the Mojave Desert or by construction workers)?

- Appropriate models

  Simplistic:

  How speculative? Identify potential damage due to credibility issues.

  Complex:

  Identify the bias: high (conservative); or low (different scenario used than plausible bounding analysis)?

- Unless one has a sense of the likelihood of the scenario, what does one do?

  — Risk management can address it.
  — Present qualitative statements about uncertainty.
  — Value of information approaches (e.g., does weather change drinking water data?).

Short-term Research:

Evaluate short-term data set: make assumptions, devise models on population variability (Ryan paper) (Wallace and Buck). Look at behavior patterns, information biases. Flesh out Chris Portier's suggestion on extrapolating 3-day data to 6 months, years. This would give the assessor some confidence in extrapolating for interindividual variability.

Long-term Research:

Collect more data. Possible ORD funding? Look at breathing rates, soil ingestion, infrequently consumed items, frequently consumed items.

APPENDIX F

PREMEETING COMMENTS

Workshop on Selecting Input Distributions for Probabilistic Assessments

Premeeting Comments

New York, New York
April 21-22, 1998

Compiled by:
Eastern Research Group, Inc.
110 Hartwell Avenue
Lexington, MA 02173

Table of Contents

Reviewer Comments

Sheila Abraham
EPA Probability Workshop

COMMENTS ON THE ISSUE PAPERS / DISCUSSION ISSUES FOR THE EPA WORKSHOP ON SELECTING INPUT DISTRIBUTIONS FOR PROBABILISTIC ASSESSMENT

Probabilistic analysis techniques are, as stated in EPA's May 1997 "Guiding Principles for Monte Carlo Analysis", viable tools in the risk assessment process provided they are supported by adequate data and credible assumptions. In this context, the risk assessor (or risk assessment reviewer) needs to be sensitive to the real-life implications on the receptors of site-specific decisions based on the analysis of variability and uncertainty. The focus should be on the site, in a holistic manner, and all components of the risk assessment should be recognized as tools and techniques used to arrive at appropriate site-specific decisions.

Preliminary (generalized) comments from a risk assessment perspective on the issue papers are provided below, as requested.

Evaluating Representativeness of Exposure Factors Data (Issue Paper #1)

1) The Issue Paper (Framework/ Checklists):

Overall, the issue paper provides a structured framework for a systematic approach for characterizing and evaluating the representativeness of exposure data. However, one of the clarifications that could be provided (in the narrative, checklists and figure) relates to the explicit delineation of the objectives of the exercise of evaluating data representativeness. The purpose of the original study should also be evaluated in the context of the population of concern. In other words, factoring the Data Quality Objectives (DQOs) and the Data Quality Assessment (DQA) premises into the process could help define decision performance requirements. It could also help to evaluate sampling design performance over a wide range of possible outcomes, and address the necessity for multi-staged assessment of representativeness. As stated in the DQA

Guidance (1997), data quality (including representativeness) is meaningful only when it relates to the intended use of the data.

On the query related to the tiered approach to ("forward") risk assessment, site-specific screening risk assessments typically tend to be deterministic and have been conducted using conservative default assumptions; the screening level tables provided by certain U.S. EPA regions have to this point also been deterministic. Therefore the utility of the checklists at this type of screening level might be extremely limited. As one progresses through increasing levels of analytical sophistication, the screening numbers generated from probabilistic assessment may require a subset of the checklists to be developed; the specificity of the checklists should be a function of the critical exposure parameters identified through a sensitivity analysis. Such analyses might also help refine the protocol (criteria and hierarchy) for assessing data set representativeness in the event of overlap of the individual, population and temporal characteristics (example, inhalation activity in elementary school students in the Columbus area exposed to contaminants at a school ballfield).

2) Sensitivity:

The utility of a sensitivity analysis cannot be overemphasized. Currently, there appears to be a tendency to use readily available software to generate these analyses; guidance on this in the context of project/ site-specific risk assessments should be provided. Providing examples as done in the Region VIII guidance on Monte Carlo simulations facilitates the process.

On the issue of representativeness in making inferences from a sample to a population and the ambiguity of the term "representative sample", process-driven selection might be appropriate for homogenous populations, but for the risk assessor, sampling that captures the characteristics of the population might be more relevant in the context of the use of the data. This issue appears to have been captured in the discussion on attempting to improve representativeness.

Empirical Distribution Functions (EDFs) versus Parametric Distributions (PDFs) (Issue Paper #2)

1) Selection of the Empirical Distribution Functions (EDF) or Parametric Distribution Function (PDF):

The focus of the issue paper is the Empirical Distribution Function (EDF), and a number of assumptions have been made to focus the discussion on EDFs. However, for a clearer understanding of the issues and to facilitate the appropriate choice of analytical approaches, a discussion of the PDF, specifically the advantages/ disadvantages and constraining situations would be beneficial. The rationale for this is that the decision on whether to apply the EDF or the PDF should not be a question of choice or even mutual exclusivity, but a sequential process that is flexible enough to evaluate the merits and demerits of both approaches in the context of the data.

In general, from a site/ project perspective, there may be definite advantages to PDFs when the data are limited, provided the fit of the theoretical distribution to the data is good, and there is a theoretical or mechanistic basis supporting the chosen parametric distribution. The advantages to the PDF approach are more fully discussed in several references (Law and Kelton 1991). These advantages need to be evaluated in a project-specific context; they could include the compact representation of observations/ data, and the capacity to extrapolate beyond the range of observed data, as well as the "smoothing out" of data. (In contrast, the disadvantages imposed by the possible distortion of information in the fitting process should not be overlooked. Further, the (traditional use of) EDFs that limit extrapolation beyond the extreme data points, perhaps underestimating the probability of an extreme event, may need to be considered. This is could be a handicap in certain situations, where the risk assessment demands an interest in outlier values. In such situations, a fuller discussion of alternate approaches such as a mixed-distribution (Brately *et al*., 1987) may be warranted.) Finally, the PDFs, given their already established theoretical basis, may lend themselves to more defensible and credible decision-making, particularly at contentious sites.

This predisposition to PDFs certainly does not preclude the evaluation of the EDF in the process. The advantage accruing from having the data "speak" to the risk assessor/ reviewer should not be minimized. Depending on the project/ site involved, the benefits of the complete representation of data, the direct information provided on the shape of the underlying distribution, and even on peculiarities such as outlier values should be discussed, as well as relevant drawbacks (sensitivity to random occurrences, potential underestimation of the probability of extreme events, perhaps cumbersome nature if the data points are individually represented). In this context, some of the comments in the "Issue/ Comments" Table ("issues" presumably derived from D'Agostino and Stephens, 1986) can serve as the basis for additional discussion.

2) Goodness of Fit:

The decision whether the data are adequately represented by a fitted theoretical distribution is an aggregative process, and goodness-of-fit is part of the sequential exercise. Preliminary assessments of the general families of distributions that appear to best match the data (based on prior knowledge and exploratory data analysis) are often conducted initially; the mechanistic process for choice of a distributional family, the discrete/continuous and bounded/ unbounded nature of the variable are evaluated. Summary statistics, including measures of shape are evaluated and the parameters of the (candidate) family are estimated. The goodness-of-fit statistics should factor into the whole process, as should graphical comparisons of the fitted and empirical distributions. Goodness-of-fit tests can be an excellent confirmatory tool for verifying the chosen distribution, when used in conjunction with statistical measures and probability plots.

However, caution should be exercised in situations where these tests could conceivably lead an analyst to support a distribution that a visual inspection of the data does not support. Also, it should be emphasized that (for example for certain physiological parameters), even if the distribution fits, maintaining the integrity of the (biological) data should override goodness-of-fit considerations. Ultimately, the persuasive power of graphical methods for assessing fit should not be underestimated.

On the question how the level of significance of the goodness-of-fit statistic should be chosen, this is often a function of the data quality assessment (DQA) for that particular site or situation; an idea of the consequences in terms of real-life examples can be gathered from EPA's Guidance for Data Quality Assessment (1997). On the whole, I tend to agree with the respondent (#4) who states that the desired level of significance should be determined prior to analyzing the data. Again, as the respondent states, if minor differences in the p-value impinge substantially on the analysis, the "conclusions are probably too evanescent to have much usefulness".

Summary statistics are useful, particularly in the initial characterization of the data (as previously mentioned). Given the constraints imposed by the project/ site logistics, all too often these are the only data available, and they have been used as the basis for analytical distribution fits (Ohio EPA, 1996). Caution should be exercised in implying a level of accuracy based on limited knowledge. Sensitivity analyses might help clarify the limitations that need to be placed in such situations particularly when dealing with an exposure parameter of considerable impact; further, the utility of such an exercise for a parameter with minor impact (as revealed by the sensitivity analysis) could be questionable.

On the question of the value of testing the fit of the more generalized distributions (presumably in lieu of the EDF), this could be an useful exercise, but the project logistics may factor into this, as also the DQA premises. Project resources available and the defensibility of the decision-making process need to be factored into the situation. The issue of fitting an artificial distribution to a data set, and ultimately arriving at a distribution removed from reality also needs to be evaluated in the project-specific context.

3) Uncertainty:

The discussion in “Development of Statistical Distributions for Exposure Factors” (Research Triangle Institute) paper is interesting in terms of the approaches suggested for evaluating parameter uncertainty; Hattis and Burnmaster’s comment cited in the paper that only a trivial proportion of the overall uncertainty may be revealed is important. Certain methods (example, bootstrapping) appear to have intriguing potential for accounting for “hot spots”.

Finally, the risk assessor/ reviewer needs to be aware that the analysis of variability and uncertainty is a simulation, based on hypothetical receptors. However, as stated initially, this sometimes academic exercise can have multi-million dollar implications, and intimately affect real-life human and ecological receptors; the risk assessor/ reviewer should always be cognizant of this consequence.

References:

Brately, P., B.L. Fox, L.E. Schrage (1987) “A Guide to Simulation”. Springer-Verlag, New York.

D’Agostino, R.B. and M.B. Stevens (1986) “Goodness of Fit Techniques”. Marcel Deker.

Law, A.M. and Kelton, W.D. (1991) “Simulation Modeling and Analysis” (Chapter 6, 325-419). McGraw-Hill, New York.

Ohio EPA (1996) “Support Document for the Development of Generic Numerical Standards and Risk Assessment Procedures”. The Voluntary Action Program, Division of Emergency and Remedial Response, Ohio EPA.

U.S. EPA (1994) “Guidance for the Data Quality Objectives Process” (EPA/QA/G4). EPA/600/R-96-055

U.S. EPA (1997) “Guidance for Data Quality Assessments - Practical Methods for Data Analysis” (EPA QA/G-9, QA-97 Version) EPA/600/R-96/084 (January 1998)

Robert J. Blaisdell, Ph.D.

Comments on Issue Paper on Evaluating Representativeness of Exposure Factors Data

The Issue Paper on Evaluating Representativeness of Exposure Factors Data is a well written, clear discussion of the theoretical issues of representativeness. I was particularly interested in the discussion of time unit differences. The Office of Environmental Health Hazard Assessment (OEHHA) is grappling with this issue with several of the distributions which we want to use for determining chronic exposure.

The issue of representativeness of a sample is often complicated by lack of knowledge about the demographics of the population under consideration. An accurate determination of the population under consideration may not be part of the risk assessment requirements of regulatory programs. If the population of concern has not been characterized, the determination of the representativeness of the data being used in the assessment is not possible.

The issue of representativeness of the sample to the population is an important question. For example, populations which are exposed to Super Fund toxicants or airborne pollution from stationary sources may be from lower socioeconomic groups. Unfortunately, most of the information which is available on mobility is from the general population. It may be that low income home owners have a much longer residency time than people of median or higher income. It may also be that low income non-home owners in certain age groups have a higher mobility than the general population. We therefore suspected that the available distributions were not representative. In addition, the U.S. Census data, the basis for the available residency distributions are not longitudinal. Another problem with the residency data when evaluating stationary sources is the issue of where the person moves to. A person moving may not necessarily move out of the isopleth of the facility. The likelihood of moving out of the isopleth of a stationary facility also may be related to socioeconomic status.

In order to address this problem, OEHHA proposed not using a distribution for residence time in our Public Review Draft Exposure Assessment and Stochastic Analysis Technical Support Document (1996). Instead we proposed doing a separate stochastic analysis scenario for 9, 30 and 70 years. We did not think that the 9, 30 or 70 years time points evaluated were necessarily representative of actual residence times, but that these were useful, reasonably spaced intervals for residents to compare with their own known residency time.

Using three scenarios complicates the analysis, but we felt that the approach had some advantages over using a distribution. The California *Hot Spots* program is a public right to know act which assesses risks of airborne pollutants from stationary sources. Public notification is required above a certain level of risk. An individual resident who has received notice is aware of the amount of the time that he or she has lived, or in many cases plans to live, in vicinity of the facility. Therefore the individual could more accurately assess his or her individual cancer risk. The relationship between the residency time assumption and the resulting risk are clear, not buried in the overall range of the uncertainty or variability of the risk estimate.

This approach might possibly be used in other cases where representative data in not available or where the representativeness is questionable. For example if the drinking water pathway is of concern and representative information is not available for the population of a Mojave Desert town, the range or point estimate of cancer risk from drinking 1, 2, 4 and 8 liters of contaminated tap water per day could be presented.

In some cases, each situation that a regulatory risk assessment program will be evaluating will be almost unique, and therefore anything other than site-specific data will not be representative. OEHHA characterized a fish consumption distribution for anglers consuming non-commercial fish using the Santa Monica Bay Seafood Consumption Study Final Report (6/94) raw data. We compared the Santa Monica Bay distribution to the fish consumption distribution for the Great Lakes (Murray and Burmaster, 1994). We found that the differences in the two distributions could be attributed to methodological differences in the two studies. Thus the assumption that a salt water fish consumption distribution was comparable to a fish consumption distribution for large fresh water body was not implausible. However, the data gathered from large bodies of water are probably not representative of small lakes and ponds with limited productivity and where other fishing options may exist. For such bodies of water a site-specific angler survey is probably the only way of obtaining representative data. For cost reasons, this option is not likely to be pursued except in a risk assessment with very high financial stakes. We chose to recommend using the Santa Monica Bay fish consumption. It could be multiplied by a fraction to be determined by expert judgment to adjust for site-specific conditions such as productivity etc. The Santa Monica Bay fish distribution may not be representative in other ways in a given situation but may still be the most practical option. It is clearly not temporally representative for chronic cancer risk assessment.

Cost is often a factor that limits representativeness.

On page 8, paragraph 3 of the Issues paper there is a discussion of determining the relationship between two populations and making adjustments in distributions based on speculative estimates of the differences in means and the coefficients of variation. Perhaps in many instances, another option would be to state that the information from a surrogate population is being used and that the actual population is known to be different, or may be different by an unknown amount. There are many questions in risk assessment for which expert opinion is no better than uninformed opinion in attempting to quantify the unknown. An example of this is the shape of the dose-response curve for cancer for most chemicals at low concentrations.. A frank admission of ignorance may be more credible than an attempted quantification of ignorance in many cases.

Robert J. Blaisdell, Ph.D.

Comments on Temporal Issues

The methods discussed for estimating intraindividual variability from data collected over varying short periods of time relative to the longer time period of interest are interesting and would appear to be useful for the NFCS data. OEHHA is giving some consideration to using the techniques described by Nusser et al. 1996 to adjust the distributions for food consumption that we have developed for food consumption using the Continuing Survey for Food Intake for Individuals 1989-91 raw data. I would be curious to know if these methods have been validated on any actual longitudinal data. The assumption of the lognormal model needed by the method of Wallace et al. (1994) may in some cases be limiting. We have discovered when we evaluated broad categories of produce consumption using the CSFII 89-91 data that some of the distributions for certain age groups were closer to a normal model than a lognormal model.

The Representativeness Issue paper discusses the importance of using current data. The continued use of the 1977-78 NFCS study is cited as an example. The raw data from the 1989-91 CSFII has been available for some time as an alternative to the 1977-78 NFCS survey. Raw data from the 1992-93 CSFII survey is now available. OEHHA has used that data to develop produce, meat and dairy products consumption distributions for the California population. It is admittedly not a trivial exercise to extract the relevant data from the huge raw CSFII data sets but this alternative has existed for several years. The 1989-91 CSFII data is clearly different in some cases from the 1977-78 NFCS. Beef consumption appears to have declined. As a matter of policy, there should be a stated preference for using the available data over attempting to use expert judgment to guess at the appropriate means, coefficients of variation and parametric model. In some of the Monte Carlo risk assessment literature, the preference appears to be for expert judgment rather than data.

The use of related data may in some cases be useful in giving some insight into the representativeness of data collected over the short term for chronic scenarios. OEHHA has used the data on total energy expenditure as measured by the doubly labeled water method to look at the representativeness of our breathing rate distribution, based in part on a one day 24 hour activity pattern survey. The information on total energy expenditure gave an indication that intraindividual variability was a huge fraction of the total variability (intraindividual plus interindividual variability).

The intraindividual variability for a broad category of produce such as leafy vegetables may not be very great relative to the interindividual variability. The intraindividual variability for a single item less frequently consumed item such as strawberries is probably much greater than for broad categories. Thus, short term survey data which looks at broader categories of produce are probably more applicable to chronic risk assessment than single item distributions.

Research Needs

The information which is needed to develop more accurate distributions for many if not most variates needed for chronic stochastic human health risk assessment are simply not available. In particular there is a lack of longitudinal data for breathing rates, soil ingestion, water consumption rates, produce ingestion, non-commercial fish consumption, dairy product consumption and meat ingestion. Some distributions in common use, such as water consumption, are based on out of date studies. More research is needed on bioconcentration and biotransfer factors. Longitudinal data on activity patterns and mobility patterns would also be very useful. There needs to be much more research on dermal absorption factors and factors which influence dermal absorption. More research needs to be done on children and the ways that they differ from adults.

Summary

The overall lack of data, particularly longitudinal data, for risk assessment variates is probably the most important single factor limiting representativeness. If the purpose of the risk assessment is to inform the exposed public, it may be possible and even preferable to use point estimates for multiple scenarios in the absence of some representative data. The statistical methods for adopting short term data for use in chronic risk assessment presented the Issue paper appear to be reasonable approaches in instances where the required data is available. More longitudinal studies would be valuable for validation of these methods as well as improving the temporal representativeness of distributions used in risk assessment. Most of the data used in stochastic risk assessment will probably be nonrepresentative in one or more of the ways discussed in the Issues paper for a long time into the future.

References

Murray DM., and Burmaster DE. (1994). Estimated distribution for average daily consumption of total and self-caught fish for adults in Michigan angler households. Risk Analysis 14, 513-519.

Nusser, S.M., Carriquiry, A. L., Dodd, D.W., and Fuller, W. A. A semiparametric transformation approach to estimating usual daily intake distributions. J. Am. Statistical Association 91: 1440-1449, 96.

Southern California Coastal Water Research Project and MBC Applied Environmental Sciences (SCCWRP and MBC). (1994). Santa Monica Bay Seafood Consumption Study. Final Report. June.

USDA (U.S. Department of Agriculture) 1989-91. Nationwide Food Consumption Survey. Continuing Survey of Food Intakes of Individuals (Data Tapes) Hyattsville, Md: Nutrition Monitoring Division, Human Nutrition Information Service.

13 April 1998

Memorandum

To: Participants, US EPA's Workshop on Selecting Input Distributions for Probabilistic Analyses

Via: Beth A. O'Connor, ERG

From: David E. Burmaster

Subject: Initial Thoughts and Comments, and Additional Topics for Discussion

Thank you for inviting me to participate in this Workshop in New York City.

Here are my initial thoughts and comments, along with suggestions for additional topics for discussion. Since I have just returned from 3 weeks of travel overseas, I will keep these brief.

1. Models and Data

In 1979, George Box wrote, "All models are wrong, but some are useful."

May I propose a new corollary for discussion? "All data are wrong, but some are useful."

2. Definitions for Variability and Uncertainty

The Issue Papers lack crisp definitions for variability and uncertainty as well as a discussion about why variability and uncertainty are important considerations in risk assessment and risk management. (See, for example, NCRP, 1996.) In particular, I recommend definitions along these lines for these two key terms:

- Variability represents true heterogeneity in the biochemistry or physiology (e.g., body weight) or behavior (e.g., time spent showering) in a population which cannot be reduced through further measurement or study (although such heterogeneity may be disaggregated into different components associated with different subgroups in the population). For example, different children in a population ingest different amounts of tap water each day. Thus variability is a fundamental property of the exposed population and or the exposure scenario(s) in the assessment. Variability in a population is best analyzed and modeled in terms of a full probability distribution, usually a first-order parametric distribution with constant parameters.

- Uncertainty represents ignorance -- or lack of perfect knowledge -- about a phenomenon for a population as a whole or for an individual in a population which may sometimes be reduced through further measurement or study. For example, although we may not know much about the issue now, we may learn more about certain people's ingestion of whole fish through suitable measurements or questionnaires. In contrast, through measurements today, we cannot now eliminate our uncertainty about the number of children who will play in a new park scheduled for construction in 2001. Thus, uncertainty is a property of the analyst performing the risk assessment. Uncertainty about the variability in a population can be well analyzed and modeled in terms of a full probability distribution, usually a second-order parametric distribution with nonconstant (distributional) parameters.

Second-order random variables (Burmaster & Wilson, 1996; references therein) provide a powerful method to quantify and propagate V and U separately.

3. Positive Incentives to Collect New Data and Develop New Methods

I urge the Agency print this Notice inside the front cover and inside the rear cover of each Issue Paper / Handbook / Guidance Manual, etc. related to probabilistic analyses -- and on the first Web page housing the electronic version of the Issue Paper / Handbook / Guidance Manual:

> This Issue Paper / Handbook / Guidance Manual contains guidelines and suggestions for use in probabilistic exposure assessments.
>
> Given the breadth and depth of probabilistic methods and statistics, and given the rapid development of new probabilistic methods, the Agency cannot list all the possible techniques that a risk assessor may use for a particular assessment.
>
> The US EPA emphatically encourages the development and application of new methods in exposure assessments and the collection of new data for exposure assessments, and nothing in this Issue Paper / Handbook / Guidance Manual can or should be construed as limiting the development or application of new methods and/or the collection of new data whose power and sophistication may rival, improve, or exceed the guidelines contained in this Issue Paper / Handbook / Guidance Manual .

4. Truncating the Tails of LogNormal Distributions

While LogNormal distributions provide excellent fits to the data for many exposure variables, e.g., body weight, skin area, drinking water ingestion rate (total and tap), showering time, and others, it is important to truncate the tails of these distributions. For example, no individual has 1 cm^2 of skin area, no individual has 10^5 cm^2 of skin area, and no individual can shower 25 hr/d.

5. Mixing Apples and Oranges

It is wholly inconsistent for the Agency to proceed with policies that legitimize the use of probabilistic techniques for exposure factors while preventing the use of probabilistic techniques in dose-response assessment. By doing so, the Agency double counts the effects of variability and uncertainty, all on a $\log_{10}$ scale -- i.e., by several orders of magnitude.

6. Report by RTI

I disagree strongly with many of the approaches and conclusions found in RTI's Final Report dated 18 March 1998.

References

Box, 1979

Box, G.E.P., 1979, Robustness is the Strategy of Scientific Model Building, in Robustness in Statistics, R.L Launer and G.N. Wilkinson, eds., Academic Press, New York, NY

Burmaster & Wilson, 1996

Burmaster, D.E. and A.M. Wilson, 1996, An Introduction to Second-Order Random Variables in Human Health Risk Assessment, Human and Ecological Risk Assessment, Volume 2, Number 4, pp 892 - 919

Bruce Hope

REPRESENTATIVENESS (Issue Paper #1)

1) The Issue Paper

We would use probabilistic methods specifically for the purpose of assessing risks from the uncontrolled release of hazardous substances at a specific location (site). Our overall goal will be to feel confident that the entire risk assessment (and not just a few of its components) is representative of site-specific conditions. Our objective is better risk management decisions. This requires us to keep a few other considerations in mind.

The issue of representativeness in terms of a fit between available exposure factors data and resulting distributions is dealt with in the issue paper. However, a risk assessment cannot be performed with exposure factor distributions alone - some type of exposure model is required. We should therefore also be concerned with the representativeness of the exposure model within which the individual exposure factors are used.

Correlation between exposure factors could significantly affect the representativeness of the resulting risk assessment. It appears possible to have too much or little correlation between factors. In some cases, the correlation is not necessarily with body weight and/or age but with an underlying activity pattern (human behavior) that may not be fully known. This nature and extent of correlation should be a factor in evaluating representativeness.

The issue of data and statistical inferences at the extreme upper bounds (*e.g.*, 99.9th percentile) of a distribution has been raised in the literature, on the Web, and in other U. S. EPA forums. As a matter of policy, we regulate at the 90th percentile, feel that decisions based on extreme upper bound estimates are potentially unreasonable, and thus have truncated the upper bound (not allowed its extension to +∞) of many of the exposure factor distributions. How any such truncation of a distribution affects its representativeness should also be discussed.

The suggestion that probabilistic methods could be used in any form of "screening-level" risk assessment is of concern. We view screening has a quick but highly conservative comparison of environmental media concentrations with published toxicity data that occurs early in a remedial investigation (RI) for the sole purposes of narrowing the focus of the baseline risk assessment. Under our current guidance, we are preserving probabilistic methods for use only in a baseline assessment.

2) Sensitivity

When various exposure factors are combined within a given exposure model, it is typically the case that a few of them have a disproportionate influence on the outcome. For example, soil ingestion rate, soil adherence factor, and exposure duration are often primary drivers, as well as major sources of uncertainty. We should broaden the discussion to consider whether all exposure factors are of equal importance, in terms of their influence on the outcome of the risk assessment, so as to better focus our distribution development efforts.

3) Adjustments

Concern has been expressed that any "default" exposure factor distributions proposed by U. S. EPA will, perhaps unintentionally, will evolve into inflexible or "standard" requirements. To counter this, as well as allow for inclusion of regional and local influences, U. S. EPA should propose, in addition to any *de facto* "default" distributions, an exemplary method(s) for establishing exposure factor distributions. This exemplary method should be as straightforward, transparent, and explainable (primarily to risk managers) as possible. It should also describe quality assurance (QA) and quality control (QC) procedures to allow for the expedient and thorough review of probabilistic risk assessments submitted to regulatory agencies by outside contractors.

EMPIRICAL DISTRIBUTION FUNCTIONS (Issue Paper #2)

{I did not have time to fully review paper #2, so only have input on this one item at this time}

2) Goodness of Fit

We should also ask, if the overall risk assessment is sensitive to both the exposure model and only a few of many exposure factors, just how "good" does every other distribution have to be in order to support credible risk management decisions? For example, if a relatively esoteric and hard to conceptualize distribution best fits available data, but a much more common and more easily understood distribution fits almost as well (say within 20%), would there not be some advantage in use of the latter? In addition, if toxicity data remain as point estimates with uncertainty approaching an order-of-magnitude, it would appear that there should be some leeway in how we choose or define certain exposure factors.

William A. Huber

Representativeness (Issue Paper #1)

1) The Issue Paper

1.1 The checklists

Section 3 of the Issue Paper regards the inferential process as consisting of several stages of inference and measurement: Population of interest -> Population(s) actually studied -> Set of individuals measured (the "sample") -> The measurements. The three stages are denoted "external" inference, "internal" inference, and measurement, respectively.

This appears to be a useful framework. However, the four checklists address the first two stages only. Checklist I concerns the "internal" inference; Checklists II through IV concern the "external" inference. No checklist specifically addresses measurement. This approach is unbalanced. The obvious parallelism among Checklists II through IV emphasizes the lack of balance. We should consider whether a better organization of checklists might be achieved. One possible organization could be:

Checklist A: Assessing measurement representativeness

Checklist B: Assessing internal representativeness

Checklist C: Assessing external representativeness

Checklist D: "Reality checks," or overview.

Checklist B and checklist I would nearly coincide. Checklist C would incorporate the (common) questions of checklists II through IV. Checklists A and D are new. Checklist A would incorporate certain questions sprinkled throughout Checklists I-IV, such as:

- Does the study appear to have and use a valid measurement protocol?
- To what degree was the study design followed during its implementation?

- What are the precision and accuracy of the measurements used in the study?
- Did the study actually measure what it claimed to?

The questions in Checklist D would focus on the fundamental questions:

- Has the data set captured the variability within the population of interest?
- Is it sufficient in size and quality to support the estimate, decisions, or actions recommended in this risk assessment?
- Can we quantify potential departures of our estimates from their correct (but unknown) values? Why and how?

Each of the bulleted items above has some detailed questions associated with it.

1.2 Tiered risk assessments

There is no subset of questions that can be selected since it cannot be foreseen which question is critical to evaluating a particular study. However, there is a basis for limiting the effort needed to establish representativeness. First, materially unimportant variables—as established, for example, by a sensitivity analysis—need not be fully addressed. Second, many of the checklist questions are relevant when variability and extreme percentiles must be characterized; they become less consequential when only a central tendency need be assessed. Finally, for a screening risk assessment, only qualitative degrees of representativeness are needed. For example, if it is known only that study results will conservatively overestimate exposures, then that study could be useful for a screening level risk assessment, but probably not for subsequent tiers.

2) Sensitivity

There are two kinds of sensitivity in a probabilistic calculation. They are related to the distinction between variability and uncertainty. We may, with some loss of generality, suppose that the calculation is a determined procedure F that processes a collection S = {p1, p2, ..., pN} of "inputs," each of which is a (possibly degenerate) probability distribution, and outputs a single probability distribution F(S). If there is a material change in inferences based on F(S) when one of the input distributions, say pI, is collapsed to a point, then the calculation is sensitive to the variability in pI. Otherwise, the distribution pI can, with some safety, be replaced by a single number (a degenerate distribution).

Uncertainty in the input pI can often be described as a collection of possible distributions {pI'} that are "close" to pI in some sense. A typical example is when pI is parametric and {pI'} is described by a set of alternate values of the parameters. There may even be a probability distribution on {pI'} (a Bayesian "prior"). If, by replacing pI by an arbitrary element of {pI'}, the inferences based on F(S) change in a material way, then the calculation is sensitive to the uncertainty in pI.

The data must be sufficient to establish either that a variable is not a sensitive input or, if it is, the data must be sufficient to characterize the variability or the uncertainty or both, depending on which contribute to the sensitivity. This provides one basis for deciding when data are adequate. However, it could be argued that any data acceptable for use in a screening risk assessment are necessarily acceptable in subsequent tiers—at a cost.

To be specific, for data to be acceptable at all they must provide some valid information about the population of interest and some quantifiable level of uncertainty must be established (no matter how great that level is). This is true for any risk assessment at any tier, not just for probabilistic risk assessments. For screening use, inputs would have to be set at extreme (but realistic) levels consistent with the data and their uncertainty, in such a way as to ensure a "conservative" estimate of risk—that is, one biased high. Once this is accomplished, it would seem there is no obstacle to using the same data in the same way in subsequent tiers, with the price for doing so being estimates that are still biased high.

3) Adjustments

Geostatistical methods are available for certain adjustments of spatial scales. Good references are Cressie, N. "Statistics for Spatial Data;" Journel, A. and C. Huijbregts, "Mining Geostatistics." In particular, methods such as "conservation of lognormality" have been developed to adjust for differences in spatial measurement scale (this has been termed the "change of support" problem). This is the spatial analog of the DW model.

Adjustments should be applied with extreme caution because results can be very sensitive to them. Similarly, surrogate data should be used very cautiously. A good point of departure for considering adjustments is the following definition, constructed to capture the use of "representative" in EPA guidance ("Guiding Principles for Monte Carlo Analysis, EPA/630/R-97/001):

> Data are "representative" when they admit objective and quantifiable statements concerning the accuracy of the relevant inferences made from them.

From this point of view, adjustments can be considered (and defended) when made in a way that allows the potential bias or imprecision thereby introduced to be quantified in the risk assessment.

EDFs (Issue Paper #2)

1) Selecting an EDF or PDF

The primary consideration is the effect the choice will have on the risk assessment results. Each choice has relative advantages and disadvantages. They come down to this: using the EDF honors the data but subjects the calculation to the risk that the EDF poorly represents population variability and percentiles, a risk that can sometimes be decreased by using a well-chosen PDF. Using a PDF requires some theory and professional judgment and subjects the calculation to the risk that either (or both) could be wrong or inapplicable.

The choice is not inherently one of preference. With small data sets especially, an EDF is unlikely to represent an upper percentile adequately and so is manifestly a bad choice. (That's not to say that any particular PDF fit to the data is necessarily better!) When measurement error is large, the EDF will not appropriately separate variability and uncertainty. On the other hand, when the data set is large and not fit well by any theoretical distribution function, using the EDF is an excellent approach.

So we come back to the basic point: what effect will choice of distribution function(s) have on the risk assessment results? This is determined in part by sensitivity analysis. For this, the exponential tail fitting approach is particularly intriguing, because it seems to provide a robust opportunity to explore how relatively more or less extrapolation beyond the sample maximum (or minimum) will influence the results.

2) Goodness of Fit

The best basis for concluding that a fitted distribution adequately represents a data set is when (1) there is a theoretical reason to presuppose the data will be represented by such a distribution and (2) the fit is consistent with that presupposition. In this situation, P-values are meaningful and useful provided that one appropriate goodness-of-fit (GOF) test is chosen before obtaining and testing the data.

Graphical examination of the distribution is crucial. All empirical distributions will depart from the theoretical fit, so the nature and amount of departure must be assessed. It is highly unlikely that any standard GOF test will produce P-values that reflect the sensitivity of the risk assessment results to these departures. In particular, goodness of fit in the upper (sometimes lower) percentiles is usually far more important than goodness of fit elsewhere.

In many cases, where many input variables are involved in a risk calculation, using fitted distributions that reproduce the means and variances of the data is likely to produce adequate results. So, more than any P-value or selection of GOF test, these three criteria will be practically useful for risk assessments:

1. Correctly represent the centers (means and medians) of the input distributions.
2. Correctly represent the variances of the input distributions.
3. Fit the important tails of the data as well as possible.

(The "important tails" are the tails most influencing the upper percentile risk estimates. The definition of the tail—e.g., data beyond what percentile—will depend on which upper percentiles are being characterized in the risk assessment.) Note that EDFs will satisfy the third criterion only when data sets are large enough to estimate extreme percentiles with confidence.

When only summary statistics are available, there is an inherent problem in fitting any distribution: it is impossible to estimate uncertainty. Using additional information about possible limits to the data (that is, what the most extreme values could be), one should over-estimate the amount of uncertainty in the fit and use that in a sensitivity analysis. Uncertainty in the variance of the data is particularly important for probabilistic risk assessments.

When the better known distributions do not fit the data, there is exceptionally little advantage to resorting to someone's system of distributions, such as the generalized F. First, there is usually no theoretical basis for adopting any of these distributions. Second, there is little assurance that the best fitting distribution in a family will adequately represent what is of importance, namely the variance and tails. Third, reproducing the calculations can be difficult if the family of distributions is not in general use or is ad-hoc, like the five-parameter generalized F distribution is. Fourth, many of these families of distributions include obscure members whose estimation theory might not be well understood or even known. It would be better for the risk assessor to work with familiar constructs whose properties (especially with regard to influencing the risk assessment outcome) are well known.

3) Uncertainty

Every standard method of assessing uncertainty has limitations. Maximum likelihood methods often are based on asymptotic normality, which sometimes is not achieved even for impractically large data sets. There are applications where the bootstrap does not work—it is not theoretically justified. Certain methods, such as pretending the likelihood function is a probability distribution, simply have no justification (based on the theory of estimation).

In general, uncertainty should be assessed as aggressively as possible. As many possible contributors to uncertainty should be considered and as many of these as possible should be incorporated in the risk assessment, because their effects accumulate.

An excellent method for assessing uncertainty is to randomly divide datasets into parts, perform calculations (such as fitting distributions, estimating statistics, and computing risk) based on each part, and evaluate the differences that arise. Certain forms of the bootstrap and its relatives, such as the jackknife, automate parts of this procedure.

Robert C. Lee, Golder Associates Inc.

Comments Regarding "Issue Paper on Evaluating Representativeness of Exposure Factors Data"

1. The issue of representativeness relates to how the risk assessor makes judgments and corrections regarding uncertainty inherent in a nonrepresentative sample. Discussion of the differences between uncertainty (bias and/or error) and variability (heterogeneity) would be useful to avoid confusion. For example, Checklist I misleadingly implies that measurement error can have an effect on variability, which is an inherent property of a population.

Uncertainty can either be characterized as systematic (bias) or nonsystematic (error). Uncertainty in exposure assessment may stem from:

Model errors

Errors in the design of the assessment method (i.e. measure of exposure)

Errors in the use of the method

Subject limitations

Analytical errors

One way to represent bias and error is as follows. A measured or observed value X_i can be represented as a function of the true value T_i, bias b, and nonsystematic error E_i, as:

$$X_i = T_i + E_i + b$$

The population distribution of Ts represents variability. However, perfect knowledge is rarely available. Therefore, E can be represented, for example, as a normal distribution with a mean of zero and variance as:

$$\sigma^2_E = \sigma^2_X - \sigma^2_T$$

where σ^2_X is the variance of the uncertain measure *X*, and σ^2_T is the true variance (assuming independence).

Bias (which can be positive or negative) can be represented as a deterministic shift in the mean of *X* as compared to the mean of *T*, as:

$$\mu_b = \mu_X - \mu_T$$

Thus, error and bias can have an effect on the estimated population distribution, but not on the true variability.

2. In many cases, an approach that uses "reference individuals" or strata rather than attempting to evaluate or estimate variability in a broad population may be useful. For instance, if one is concerned about children's exposure to lead in a Western mining town, it may be simpler as a first step to hypothesize a few examples of children with deterministic characteristics with regard to site-specific population variability, and then evaluate the uncertainty associated with these reference individuals' exposures. This method can be relatively inexpensive and easy compared to population sampling, and could be used as a screening step in an iterative decision-making framework.

3. The exact meanings of the terms "probability sample" and "probability sampling" as used in the issue paper are unclear. Presumably these are broad terms covering schemes such as random, stratified, cluster, composite, etc. sampling. If so, then there should be clarification and discussion regarding the methodological and inferential differences between these methods. For example, simple random sampling may not be appropriate for all environmental exposure variables. If an exposure factor varies geographically, then it may be more appropriate to spatially stratify the population, and characterize the factor within each strata as accurately and precisely as possible.

4. As stated in the text (page 8, final paragraph), the process of determining the "importance of discrepancies and making adjustments" may be highly "subjective". However, the remainder of the discussion focuses heavily on frequentist methods of accounting for sources of uncertainty, which may not be the most appropriate approach. There should be discussion regarding both empirical and nonempirical Bayesian methods of population inference, since these methods are very powerful and are increasingly used in risk applications. A major advantage of Bayesian methods is that they allow refinement or "updating" of *a priori* knowledge with additional data or information.

5. More attention is devoted to "temporal" characteristics of a population than "individual" or "spatial" characteristics in the text. The reason for this is unclear. There should be discussion of how to determine the relative importance of these characteristics in risk assessment.

6. Discussion of Bayesian techniques may be useful in Section 5 of the paper, which covers issues involved with improving representativeness.

7. Discussion of the use of simulations for future scenarios would be useful. For example, if a the characteristics of a population are changing over time, time trends could be incorporated into a simulation to determine the parameters of an particular exposure variable in, say, 20 years.

Comments Regarding "Issue Paper on Empirical Distribution Functions and Nonparametric Simulation"

Robert C. Lee, Golder Associates Inc.

1. The assumptions listed in the Introduction of the Issue Paper are important and should be discussed further. The first assumption,". . .data are sufficiently representative of the exposure factor in question", is rarely met. Uncertainty associated with representativeness is often considerable. The second assumption, ". . .the analysis involves and exposure/risk model which includes additional exposure factors", is often true, although evaluation of the upper tail of a variability distribution is often difficult because of its uncertainty. If the tail is of interest, it may be preferable to stratify the analysis so that the mean of a high-exposure stratum can be used in the risk assessment. The third assumption, ". . .Monte Carlo methods will be used to investigate the variation in exposure/risk", may be true in practice, but other simple analytical and numerical methods exist. Given simple distributional assumptions (e.g. lognormality), a hand calculator can be used to calculate probabilistic output of many regulatory risk assessment models.

2. Examples of EDFs that have been used in risk assessments would be useful.

3. The statement implying that it is rare that theoretical probability distribution functions are "available" for exposure factors deserves discussion. For example, under the maximum-entropy criterion, theoretical PDFs may be fit in a rigorous manner using various combinations of limited *a priori* information. Furthermore, the assumption of lognormality for many exposure variables and models has a theoretical as well as a mechanistic basis. It is hard to argue against using lognormal distributions when non-negative, unimodal, positively skewed data are available.

Regardless, there is a practical continuum between using an EDF and, say, a maximum-entropy theoretical distribution. The issue of sensitivity is important; i.e. when does it make a difference in a risk assessment? In general, EDFs may take more time to develop. Discussions of the utility of particular distributions should be separated from theoretical arguments. An iterative approach to refinement of environmental exposure distribution functions should be discussed. This could potentially avoid inefficiency, and could be used to focus research dollars. If conducted within a Bayesian framework, prior EDFs or PDFs can be refined given additional data.

4. Much discussion in the text centers on the appropriateness of particular goodness-of-fit methods, visualization, etc. All of these methods are "blunt tools". Most statisticians simply use a number of different methods simultaneously or iteratively. If all the methods agree that a particular parametric distribution "fits" the data, then that distribution is probably appropriate. If they disagree, then the mechanistic and statistical justification for a particular distribution form and the sensitivity of the model output to the distribution defined should be examined; an EDF may be more appropriate. If the model output is insensitive to the particular PDF defined for a particular variable, then it probably does not matter what shape it takes.

Samuel Morris

Comments on Issue Paper on Evaluating Representativeness of Exposure Factors Data

3.1 Inferences from a sample to a population

The population of concern at a Superfund site is generally the population surrounding the site. This is true if the concern is for exposures during remediation activities. If there is some residual risk that may last over an extended time, the population of concern may change. In a brownfields situation, for example, the population of concern may be people who will work at the site years into the future. These people may be quite different than the population currently living around the site.

4. COMPONENTS OF REPRESENTATIVENESS

There is no question that one would like a clear definition of the population of concern, but if a representative sampling of the characteristics of that population has not been done, that definition doesn't exist. Isn't that why one uses information from a surrogate population? That question then is, if one cannot characterize the population of concern, how can one know if the surrogate population is suitable to represent the population of concern? The answer is a practical one. It depends on the availability of resources, which in turn one hopes depends on how severe the risk is judged to be.

4.1 Internal components - surrogate data versus the study population

Certainly the representativeness of the surrogate study for its own study population should be evaluated. This paragraph seems to suggest that every assessor that makes use of a surrogate study should make this evaluation. Good surrogate studies are generally used over and over again by many assessors. Such an evaluation should only need to be made once, with the results made available to all assessors. Along with this evaluation should be an evaluation of the character of the population for which the particular surrogate study is useful. This could go further to provide some limiting population characteristics beyond which the surrogate would not be recommended.

4.2 External components - population of concern versus surrogate population

The suggestion of using several national Food Consumption Surveys as a basis to extrapolate dietary habits into the present or future seems like a rather precarious thing to do. It also is something that could only be done for an extremely large, important, and well-funded assessment. It is another study that, if done at all, should only be done once and results made available widely.

Regarding several assessors independently speculating on the mean and coefficient of variation of a parameter (expert judgment?), to avoid the phenomenon of anchoring, a useful protocol is to have the experts begin from the extremes and probabilities toward the central point, rather than beginning with the mean.

Checklist I.

I don't understand the questions, "For what population or subpopulation size was the sample size adequate for estimating measures of central tendency . . .and other types of parameters?" The previous questions ask if the sample size was adequate, etc. Presumably this means it is adequate for the size of the population that was studied. I am assuming that this checklist pertains to an internal analysis of the surrogate study and has nothing at this point to do with a different population that is of concern to the assessor.

Checklist II.

I suspect that in most situations, the answer to the first question will be that the two populations are disjoint.

Checklist III.

These questions concern whether the two populations inhabit the same geographic area. Presumable the interest is in similar climate, activity patterns, etc. Spatial characteristics convey a broader–in fact a different–meaning to me. It suggests how the population is distributed in space. Is it a high density area or a low density area? Are there clusters of housing separated by open space?

Responses to the Questions on Representativeness

Issue Paper on Empirical Distribution Functions and Non-Parametric Simulation

Introduction

Is stochastic variability really the right term here? Just to make sure I am interpreting this right, I take "variability" to mean that, for example, some people drink more tap water than others and thus have a greater exposure. The big difference between variability and scientific uncertainty or random error is that it is presumably possible to identify which individuals drink 2 liters/day and which drink 0.5 liters/day, or they can identify themselves. This is important because it provides a tool for intervention. For example, we can warn pregnant women to reduce their intake of fish rather than setting a standard requiring everyone to eat fewer fish. "Stochastic variability" seems to imply variability that is so randomized that we–nor the individuals involved–cannot determine who has a high exposure and who has a low exposure. In that sense, it is the same as a cancer dose-response function.

Why do we write-off the use of theoretically based distribution functions? Many environmental variables do seem to be distributed lognormally. It isn't just coincidence. I believe that we are often better off fitting our data to a lognormal than trying to develop an empirical distribution based on what is typically a rather small data set. I once got some good advice when I was a junior engineer trying to figure out how much water was flowing in a pipe. My boss told me, "We have a good theory explaining the flow of water in pipes, but our meters have a 5% error at best. If there is a difference between the theory and the data, assume the meters are wrong." My only problem with lognormals is how well they continue to map nature out in the extreme tails. Even there, however, how much confidence do we have in the 99th percentile of an empirically based distribution?

Part 1. Empirical Distribution Factors

Extended EDF

The EDF is extended by adding plausible lower and upper bounds, but the paper does not mention how one extends the linearized curve to reach those bounds. Presumable by using a curve-fitting routine of some kind.

In many cases, there is no clearly obvious point for the upper or lower bound. We know we do not have any one kg adult males, but how do we decide to stop at 15 kg and not 14? Expert judgment is used. Expert judgment may be all we have, but it is not a great justification, and it is important that we provide justification. I believe it is worthwhile to do a sensitivity analysis to find the difference between using quasi-arbitrary bounds and letting the curve run out to zero or infinity. It might also be worthwhile to check the difference with stricter, but perhaps more reasonable bounds, say a 40 kg adult male.

Samuel Morris

Mixed Empirical-Exponential Distribution

I think that mixing theoretical distributions with empirical distributions in some kind of composite sounds like a good idea.

Starting Points

The smaller the data set, the greater the rationale for using a standard distribution.

Responding to #5, people feel more comfortable with a theoretical distribution because it has a theoretical basis that supports interpolation between data points and extensions beyond the data, although I was always told never to do the latter. When plotting empirical data without a theory, one never knows if there is some big discontinuity between two completely innocent looking data points. The problem is that the theory behind the distribution is mathematical, not physical. To be comfortable with interpolating or extrapolating in either case, one must have a theory of the physical process involved.

P. Barry Ryan

Workshop on Selecting Input Distributions for Probabilistic Assessment

In the transmittal letter dated March 27, 1998, Beth O'Connor asked us as reviewers to provide "... not... comprehensive comments, but rather your initial reaction and feedback on the issues... ." Further, we have been asked to focus on the so-called "Representativeness" Issue Paper. My discussion focuses on that manuscript to start.

First Reactions

My first thoughts on this paper center on the need for an "audience" to be selected. Issue papers such as this one will lead, eventually, to guidance documents similar to those supplied as background reading. But what is the audience of this document? To a degree, the audience must be viewed as one and the same. This document will be referenced in a guidance document. Assuming this, a diligent worker looking for more information will seek out this manuscript. Hence it should be readable and accessible to practitioners of risk assessments and exposure assessment science. With this assumed audience in mind, I continue with my initial reaction to the Issue Paper.

The **Introduction** commences with a single sentence that concisely described the purpose of the document. This is a good start; the reader is entitled to know what is being discussed. Unfortunately, the next sentence is a parenthetical notation. Is this statement unimportant, less important, to be ignored, or what? The third sentence has a relative pronoun as the first word but the antecedent is unclear. To what does "This" refer? Exposure factors? Representativeness? Whatever it may be, it is both extremely brad and extremely important as the rest of the sentence tells us.

Before the above is dismissed as grammatical nitpicking consider the following. At this point, we are only three sentences into the document and I, considered to be an expert reviewer, am uncertain as to what is being discussed. A gentle introduction to a difficult subject goes a long way toward keeping the reader "on line." A little editing for style up front will make this document much more useful.

Let us continue. The next paragraph is a roadmap describing the way through the remainder of the document. These two paragraphs provide the **Introduction**. More is needed. Why is this important? When should it be applied? What has been done in the past? These are all reasonable questions to ask.

The next section begins the meat of the Issue Paper. **General Definitions/Notions of Representativeness** is a real mouthful of a title. The term "Notions" has the connotation of uncertain knowledge. Definitions are quite the opposite. Will we be treated to contradictory information in this section? Apparently the answer is "Yes" because, as pointed out the Issue Paper continues, a reference to Kruskal and Mosteller indicates that the term on which we are seeking guidance has no "... unambiguous definition..." Why is it necessary so early on in the discussion to confuse the issue in the mind of the reader by saying that no definition exists? Why would a reader of this document continue reading rather than throwing his or her hands up in despair?

The next paragraph (and accompanying table) adds further fuel to the fire. What is the purpose of this table? How does it contribute to the definitions or notions of representativeness? There is no discussion of the importance of the terms, how they might be used in assessing representativeness, nor the purpose of the table.

So, again, we have a section that needs significant editing. It is not clear to me that this section adds any insight into the notion (or definition) of representativeness. Th elementary concept is not difficult. The attempt to be all-inclusive at the very beginning, however, is doomed to failure. It is difficult to tell someone what works by telling him or her all of the problems with the system first. It would be better to adopt a working definition, show how it can be applied to many situations, then list some problems with the working definition. This allows the reader to gain some understanding of the concepts, without having to grasp the entire subject *a priori*.

I have, until this point, spent a great deal of time discussing a very small part of the Issue paper. In particular, I may have spent more space on the discussion than the manuscript length to this point. However, the first page or two of any document sets the tone for the whole piece. The tone for this manuscript ranges from one of despair to one of disorganization. There is very little room in that continuum for gaining new insight. I urge a re-write of these early sections.

Moving on to the next section, **A General Framework for Making Inferences**, begins the "meat" of the manuscript. As a matter of style, I do not care for a series of parenthetical notations in sentences. I believe that it obscures the meaning of the prose. Shorter sentences fully describing each of the activities are better. This is a recurring style point throughout the document. I will not comment on it further.

Figure 1 represents a nice, concise "decision tree" approach to risk assessment data collection. The discussion is muddied somewhat by the introduction of the (undefined) concept of surrogate data. Reordering of sentences in the paragraph to bring the example closer to the first use of the word surrogate would clarify substantially. But we quickly go far afield from our discussion of representativeness. The manuscript needs to focus on this concept. Indeed, the entire section on Inferences seems misplaced. Should it not be at the end of the document? On the other hand, Figure 1 *is* useful to the discussion of representativeness. The branches in which one must assess this factor offer an excellent opportunity to introduce techniques, etc., to assess representativeness. For example, the figure instructs the reader to follow the algorithms outlined in checklists I-IV. Why not discuss them now? It would seem that a discussion of Figure 1 in light of representativeness would be a more useful first step than to develop concepts of inference form it. The figure is designed to result in an inference, granted, but the pedagogical role of the figure here is to help the reader understand the concept of representativeness.

The next section, **Components of Representativeness**, begins to dissect the concept into pieces more manageable. The table, Table 1, and the coupling of the discussion to the Checklists in the appendix, are perhaps the strongest parts of the Issue Paper. Table 1 is especially noteworthy. It presents the fundamental questions and parses them out according to the "population" characteristics under investigation.

These include Individual Characteristics, Spatial (here misspelled as "Spacial") Characteristics, and Temporal characteristics. Further, the characteristics are divided between exogenous and endogenous effects- a very useful division. The focus should remain on this table. Discussion should expand, examples given, and understanding reached. These are the essential concept of the Issue Paper.

Unfortunately, the manuscript gets bogged down a bit at this point with the "Case" scenarios. I kept getting confused between Case 2, Case 2a, etc. Also, the introduction of the National Food Consumption Survey confused rather than helped. I found myself wondering if this approach was only applicable to the NFCS or did it have more general applicability. The topic is very general and the specificity of the example obscured that. Again, the tabular presentation is much more straightforward and helpful. Table 2 could be discussed without reference to the NFCS and the different components of representativeness addressed much more clearly and generally.

With section 5, **Attempting to Improve Representativeness**, the tenor of the Issue Paper changes dramatically to become much more statistical in nature. It also becomes more difficult to follow. At points in this section, the authors go off on tangents. See for example the discussion on raking techniques on page 12. A better approach would include more on when such data are likely to be suspect and a better description of the weighting techniques that have been advocated.

In the sub-section **Adjustments to Account for Time-Unit Differences**, there is considerable discussion of the Wallace, et al., approach to inferring temporal effects. No mention is made, however, of the work of Slob (See Risk Analysis **16**, 195-200, 1996) who advocates a different technique and evaluates both. Regardless of this missing reference, one questions why it is here at all. It is very detailed and, in my opinion, should be described briefly in terms of its logic, then detailed in an Appendix. The brief reviews of the Clayton et al., paper, the two Buck, et al., papers, the work by Carriquay and co-workers should receive the same treatment.

The section **Summary and Conclusions**, is really only a summary. The first two paragraphs perhaps should have come earlier in the document rather than at the very end. They express the philosophy of what needs to be done. This is a good thing- it sets the stage for the Issue Paper.

Continued Thoughts

After the above impressions while reading the document, I have come away with the impression of a fairly uneven presentation that may not be especially valuable either to the risk assessment community nor to EPA. The idea of an Issue Paper addressing the concept of representativeness is a good one. Data are often used in a willy-nilly fashion with little regard for the way in which they were collected not what the study design intended to do. Because of this, erroneous conclusions can be drawn resulting in much wasted effort and, sometimes, money.

I think the document as now presented does not present the issues well. However, the Figure, Tables, and Checklists are excellent. They provide a strong foundation for a document useful for both the neophyte and expert alike. As an exposure assessor, I am always trying to come up with clean definitions of the parameters I am measuring. IS it exposure? Is it dose" Is it applied dose? The authors of this Issue Paper draft have crafted answers to similar questions associated with the representativeness of data, surrogates for data, and the pitfalls of ignoring the problem altogether. Unfortunately, these gems are buried in a veritable rockslide of other information. They are not given their proper attention in the Issue Paper. The science and EPA would be well served by asking for a re-write based on the Figure, Tables, and Checklists. Some introductory prose should be placed up front to set the stage- perhaps the two paragraphs (or modifications thereof) found at the beginning of the Summary. This material would be descriptive of the problem at hand answering questions such as why is representativeness critical, how is it often lacking, and why attempts to improve the representativeness of sample must be done carefully. This would then be followed by Figure 1 and its description, which leads further on to Table 1. The description of Table 1 and Figure 1 give the essentials of the representativeness argument.

The next section would use Table 2 as its focus. Table 2 expands on the ideas of Table 1 and thus is an excellent follow on. The "examples" could be relegated to an appendix with more complete examples chosen and more detailed calculations worked out.

Finally, the Checklists should be given a more prominent placement, and a more complete discussion.

Mitchell Small

Comments on Pre Workshop Issue Papers:

"Evaluating Representativeness of Exposure Factors Data," and "Empirical Distribution Functions and Non-parametric Simulation" for US EPA Workshop on Selecting Input Distributions for Probabilistic Assessment

(New York, NY; April 21-22, 1998)

Issue Paper on Representativeness

Overall

I find the discussion of representativeness in this first issue paper to be generally thoughtful and helpful. The paper does a good job of presenting statistical concepts of experimental design in a manner that should be understandable to most exposure modelers. The major issues of target populations versus sampled or surrogate populations, and differences in available vs. desired spatial and/or temporal coverage and scale, are addressed in a clear and comprehensive manner.

Tiered Approach and Sensitivity Analysis

The issue of tailoring the framework to a tiered approach to risk assessment is integrally linked to the importance and need for sensitivity analysis when the tiered approach is used. When simpler screening level assessments are pursued, sensitivity analysis is critical to determine whether a significant problem, worthy of attention or remediation, could occur. Sensitivity analysis is always most meaningful in a decision analytic framework - can the decision derived from the risk assessment change as a result of a change in the simplifying assumption (in this case, the use of data or distributions derived from a sample of questionable representativeness)? The only way to determine whether this is so is to repeat the analysis with the underlying data or derived distributions modified in a manner consistent with known or suspected differences, over the range of plausible adjustments.

If a plausible adjustment does lead to a change in the risk management decision, then the analyst must first consider a more rigorous basis for determining the adjustment. If, with a better basis for making the adjustment, the range of predicted exposure or risk still "straddles" multiple decisions regimes (i.e., different management decisions are still possible given the improved adjustment and the overall uncertainty from other assumptions/parameters in the assessment), then this suggests the need to move to the next level of sophistication in the tiered approach. This could include the use of a more detailed and rigorous exposure and risk assessment model, as well as collection of a more representative sample for the target population.

Adjustment

The discussion of methods for modifying statistical estimates derived from a surrogate population to obtain results applicable to a different target population is thorough and informative. I do have a few insights to add on encouraging the use of hierarchical models with covariates to derive more representative distributions for the target population; on variance adjustment methods for spatial data; and on the use of Bayesian methods for combining information from surrogate (e.g., national) and target (e.g., site-specific) samples.

<u>Adjustments based on covariates:</u> The discussion in Section 5.1 covers the usual methods for weighting sample observations or sample statistics to adjust for stratification of the target population in the sampled population (either intended, as is the case in a pre-planned survey of the target population, or unintended, as is case addressed in the issue paper, when the stratification weights are a matter of happenstance). The discussion does recognize the utility of covariates (either continuous or discrete) for determining sample weights and mentions the method of "raking" for deriving these.

I think more could be done to encourage the collection and use of covariate data, in particular, using these data to develop "derived distributions" for the target population. Derived distributions arise when a relationship between the parameter of interest and the covariates can be established in a surrogate population. [This relationship could be modified for the target population based on a small sample and Bayesian methods (see my discussion below for how this might be done).] The relationship is combined with the distribution of the covariates in the target population to derive the distribution of the parameter of interest in the target population. The relationship need not be deterministic -the method is quite amenable to use with the usual regression relationships (with explicit distributions of residuals) that are developed in exposure assessment.

Consider the following example with a simple, closed-form solution: For subgroup j (i.e, based on gender, ethnicity, urban vs rural, etc.), the natural logarithm of house-dust lead, ln(house-dust lead), for person k is related to income, I, with the following relationship:

$$\ln(HDL_{k,j}) = a_j + bj[\ln(I_{k,j})] + e_{k,j}$$

where a_j is the intercept, b_j the slope and $e_{k,j}$ the residual of the regression relationship, with

$$e_{k,j} \sim N(0,\sigma_{ej}).$$

If income I for subgroup j is lognormal:

$$\ln(I_{k,j}) \sim N(\xi_{ij}, \phi_{Ij})$$

then HDL for subgroup j is also lognormal with

$$\ln(HDL_k) \sim N(\xi_{HDLj} = a_j + b_j\xi_{Ij}\,,\ \phi_{HDLj} = [b_j^2\phi_{Ij}^2 + \sigma_{ej}^2]^{0.5})$$

The distribution of HDL for the entire target population with subgroup proportions P_j, is the P_j-weighted mixture of the lognormal distributions determined for each subgroup.

For more complicated relationships between the parameter of interest and the covariates, or a more complicated distribution of covariates in the target population, Monte Carlo simulation methods may be required to derive the distribution. An example of this (entitled, "Bayesian Analysis of Variability and Uncertainty of Arsenic Concentrations in U.S. Public Water Supplies," by Lockwood, et. al.) is attached. It presents early results of a project for the EPA Office of Ground Water and Drinking Water (OGWDW) to estimate a national distribution of arsenic occurrence in source water used by drinking water utilities, based on a stratified national survey. The application is an example of Case 3 in Table 2, where the surrogate population is a subset of the population of concern. The most pertinent part of the attachment is highlighted, noting that the national distribution is synthesized by sampling the covariates of the target population.

The use of covariates for deriving distributions of exposure factors in a target population is a powerful tool that should be encouraged in the issues paper with more examples and methods. It would also encourage exposure assessors and analysts to be more careful and thorough in their collection of covariate data as part of their monitoring programs.

Variance adjustment for spatial data: The report does a good job covering the options for adjusting bias and variance for time-unit differences; similar methods can be utilized for differing scales of spatial representation. A good reference for this is Random Functions and Hydrology (Bras, R.L. and I. Rodriguez-Iturbe, Addison Wesley, Reading, PA, 1985), especially Section 6.8, Sampling of Hydrologic Random Fields. Methods are presented for accounting for spatial correlation when determining the variance of an area average. (The other thing we should do is vote on the correct spelling of spatial/spacial.) Bayesian methods for combining information from surrogate- and target-population samples: I have learned a lot recently about Bayesian methods for combining expert judgment and observed data to estimate distributions. Some of these are discussed in the attached paper by Lockwood et al. The Bayesian method allows a prior judgment for distribution parameters to be updated based on an observed data set, yielding a posterior distribution for the distribution parameters. The posterior distribution characterizes the uncertainty in the resulting estimation, but can also be used for "best-fit" point estimates (e.g, based on the mean or mode of the posterior distribution). Bayesian estimates converge to those of classical methods when "vague" or "informationless" priors are used, so that the information in the sample dominates that of the prior.

Bayesian methods can add a lot to the suite of tools available for using surrogate population samples when estimating target population statistics. A number of these tools are described in a paper that Lara Wolfson and I are (hopefully!) about to complete, "Methods for Characterizing Variability and Uncertainty: Bayesian Approaches and Insights" (we have been "about to finish this paper" for quite a long time, covering a few of our recent meetings - hopefully I will bring a copy to the meeting in New York). In particular, estimates from surrogate population samples can serve as priors for the target population, allowing information from (presumably small and limited) site-specific studies to be informed by, and combined with, the previous studies of the surrogate population. Results from multiple surrogate populations can also be used, each given a weight, along with the informationless prior, to determine how much the resulting estimate will be based on each of the surrogate population studies vs. the information in the target population survey itself.

A Specific Comment on the Representativeness Paper:

The discussion on "Summary Statistics Available" in Section 5.1 (page 10) contains what I believe to be an error, when suggesting that standard deviations be averaged across subgroups when approximating a population standard deviation: "In the case of population variance, we recommend calculating the weighted average of the group standard deviations, rather than their variances, and then squaring the estimated population of concern standard deviation to get the estimated population of concern variance." However, neither of these approaches properly accounts for possible differences in the means across the subgroups, which also contribute to the population variance. The correct approach is to compute E[X2] for each subgroup:

$$E[X_g^2] = E^2[X_g] + Var[X_g]$$

then $E[X^2]$ for the population:

$$E[X_{ATP}^2] = \Sigma_g P_g E[X_g^2]$$

and finally, the variance of X for the population:

$$Var[X_{ATP}] = E[X_{ATP}^2] - E^2[X_{ATP}]$$

where $E[X_{ATP}]$ is computed using the middle equation on page 10.

Issue Paper on Empirical Distribution Functions and Non-parametric Simulation

You appear to have already gathered a lot of thoughtful comments on the two topics addressed in this issue paper. Will any of these respondents be at our meeting? Will they be identified? I have given more thought to Part II (Issues related to fitting theoretical distributions) than I have to Part I (Empirical distribution functions). I identified strongly with the comments of Respondent #6 in Part II. To add slightly to Respondent 6's comments, I note that parametric tests of significance for the fit of a TDF almost always reject a particular parametric form as the sample size gets large - real populations invariably exhibit some deviation from a theoretical model, which cannot capture all of the population's behavior and nuances. In these cases, visual comparisons of observed and fitted distributions are essential for determining whether these deviations are in fact important to the problem at hand.

Review Comments on "Issue Paper on Evaluating Representativeness of Exposure Factors Data."

(March 4, 1998 Report)

by Edward J. Stanek III

Are questions of differences in populations, questions related to differences in spatial coverage and scale, and questions related to differences in temporal scale complete? Should other areas be added?

The document defines a population in terms of a set of units (subjects) at a location and time, a definition that is a standard starting point for traditional survey sampling. The definition of the population is important, since the term "representativeness" is being used to describe the relationship between estimates of exposure, and the true exposure of subjects in the population (or summary measures of these true exposures). An example of a typical population is (p3) "the population surrounding a Superfund site".

The population is defined as a "snapshot" of persons in time and space. Although this definition fits the traditional survey sampling paradigm, this definition may be lacking from the stand point of defining exposure in the context of the public's health. The photographic like quality of the definition does not account for the fact that new people may move into the picture, and others may leave after a short time has passed. Thus, while "representativeness" may be assessed for the picture, the picture itself may be limited. As a result, the assessment of representativeness may have limited relevance for exposure and ultimately the public's health. Of course, when one looks at the "snapshot" close to the time it was taken, the differences may be slight. After a longer time period, the differences may be dramatic. This practical concern over defining the "population" is ignored in the report.

It is important to introduce a longer time frame and possible changes in exposure when defining exposure in a population. Such definitions are important conceptually, pragmatically, and politically since they define the target parameters for exposure. Such definitions are accessible to a broad range of interested parties and not limited to statistical or technical experts. They set the stage for decisions on additional data collection, and technical choices for estimation and modeling. The current document limits the scope of "representativeness" by defining it only in a context that has an established traditional statistical literature.

In a simple sense, such a definition may be diagramed as in Table 1. The idea is that over chronologic time, there will be mobility and other physical changes. Thus, exposure for the first subject (ID=1) may differ between 1998 (E_{11}) and 1999 (E_{12}). Similarly, ID=1 may move in the year 2000, and hence no longer be exposed. Other subjects may move in the area. Subjects will also age, and their exposure may change with age. Of course, the exposure values in Table 1, while potentially observable, are not known. Nevertheless, a consensus on what will constitute such a potentially observable exposure table is the starting point for discussion of "representativeness". This conceptual framework has a rich background (Little and Rubin (1987)).

The present document defines the problem in terms of the shaded cells in Table 1. I suggest that the starting point should more closely correspond to a population as defined in Table 1. Establishing the goal first will help prioritize issues such as representativeness, sensitivity, and adjustments. One might dispute this goal by arguing that the problem definition is difficult, exceedingly complex, and since conceptual, detracts valuable time and effort from what data is known. I would argue that establishing consensus on this definition (while not statistical) should be the starting point for "evaluating representativeness". The definition itself is likely to force an expansion of the context to more modern sampling literature, such as super-population models and methods and model-based inference (Cassel et. al., (1977), Scott and Smith (1969), Meeden and Ghosh (1997)).

Table 1. Potentially Observable Exposure on Subjects in the Defined Spatial Location (E_{ij} =Exposed)

Time (Yr) (j)	Subject IDs (I)								Average
	ID=1	ID=2	ID=3	...	ID=N	ID=N+1	ID=N+2	ID=N+3	
1998	E_{11}	E_{21}	E_{31}	...	E_{N1}				μ_{1998}
1999	E_{12}		E_{32}	...	E_{N2}	$E_{N+1, 2}$			μ_{1999}
2000			E_{33}	...	E_{N3}		$E_{N+2,3}$	$E_{N+3,3}$	μ_{2000}
...									
	μ_1	μ_2	μ_3	...	μ_N	μ_{N+1}	μ_{N+2}	μ_{N+3}	μ

Are there ways of formulating questions that will allow a tiered approach to risk assessment (a progression from simpler screening level assessments to more complex assessments)?

A general strategy for tiering estimation approaches is by ordering the assumptions. With very extensive assumptions, all exposure assessments are easy. For example, assume that everyone at every time in every location has the exact same exposure, and that this exposure can be measured without error. Using these assumptions, a single measure on a single subject will suffice. These assumptions are clearly too strong to be broadly acceptable. Nevertheless, these assumptions represent an extreme which has as an opposite extreme the target "potentially observable" population (which is exceedingly complex). A gradation of assumptions can be formed between the two extremes, with such a framework leading to a tiered approach.

The framework asks how important are (or sensitive is the analysis to) population, spatial, and temporal differences between the sample (for which you have data) and the population of interest. What guidance can be provided to help answer these questions?

The document addresses the way the "surrogate population" represents the population, how the sample from the surrogate population relates to the surrogate population, and finally, how the measured value relates to the true value for the measured unit. Assuming that the population defined is the potentially observable population of interest, this is a good framework for developing inference. Some guidance can be provided to structurally evaluate the sensitivity of the exposure estimates to analysis decisions. To do so, we build estimates from the data to the surrogate population, and finally to the population.

Table 2 represents a framework for successive development of estimates to the population. Probability sampling will connect the surrogate data to the surrogate population, and may serve as the basis for inference to the lower shaded portion of the Surrogate Population. Specifically, the inference consists of estimates of population parameters, and the accuracy (mean squared error) of those estimates. Non-response, limited coverage, etc. may require additional assumptions before inference can be extended to the entire Surrogate Population.

Improvements in the accuracy of estimates for the surrogate population may be possible via modeling and/or post-stratification. The models developed on surrogate data may provide support and serve as a structure for assumptions needed to predict exposure in the surrogate population not stemming from the probability sample. For example, models based on surrogate data may develop a strong dependency of exposure on age and gender, but a weak to null relationship with urban /rural geographic location in one state. Assumptions to estimate exposure in another state (the portion of the Surrogate Population requiring assumptions) may be supported by evidence from the surrogate data, although not directly linked by the probability sampling inferential framework. The range of sensitivity analysis (for example, varying the urban/rural exposure relationship) can be established making used of model based estimates when extending inference to the non-sampled surrogate population.

Models and assumptions most likely will be the primary source to generate estimates from the surrogate population to the population of interest. As the distance increases from the actual data, the role of the models and assumptions will increase. This increased role will result in the estimates being more sensitive to the assumptions. Much progress is currently being made in studying issues of sensitivity similar to these issues in epidemiology, where a similar situation occurs in observational epidemiologic studies (see recent presentations by Wasserman, Rotnitzky et al. (1998)). Three-dimensional sensitivity plots, such as those developed by Rotnitzky et al., provide a way of visually communicating and identifying the relative importance of assumptions .

Table 2. Conceptual Steps in Developing Inference from Data to the Surrogate Population to the Population of Interest.

		Population of Interest (Assumptions Required)
	Surrogate population (Assumptions Required)	**Population of Interest**
Data From Surrogate Population	**Surrogate Population**	

Adjustments

The description of adjustments focus on adjustments due to time unit differences. There are empirical ways of dampening short time variation when estimating longer time interval distributions that do not require parametric assumptions (such as the log-normal assumptions illustrated by Wallace et al (1994). Such methods (such as empirical Bayes methods) require some assumptions, but the assumptions may be minimal and subject to verification. More research is clearly needed in these areas. This is however an active research area that is close to providing answers to practical concerns.

References

Cassel, C-M, Sarndal, C-E. And Wretman, J.H. (1977). Foundations of inference in survey sampling. John Wiley and Sons, New York.

Little, R.J.A., and Rubin, D.B. (1987). Statistical analysis with missing data. John Wiley and Sons, New York.

Meeden, and Ghosh, M. (1997). Bayesian methods for finite population sampling. Monograph on Statistics and Applied Probability, Chapman and Hall, New York.

Scott, A.J. and Smith, T.M.F. (1969). "Estimation in multistage surveys," Journal of the American Statistical Association, 64:830-840.

Wasserman, L. (1998). A tutorial on G-estimation. Eastern Regional North American Annual Meeting of the Biometrics Society, Pittsburg, PA., April 1, 1998.

Rotnitzky A. , Robins, J., and Scharfstein, D. (1998). A G-estimation approach for conducting sensitivity analysis to informative drop-out and non-ignorable non-complacence in a randomized follow-up study. Eastern Regional North American Annual Meeting of the Biometrics Society, Pittsburg, PA., April 1, 1998.

Response to Questions on Issues Paper #1 (Representativeness)

Alan Stern, Dr.P.H., DABT
Div. of Science and Research
New Jersey Dept. of Environmental Protection

I believe that the "checklists" are a conceptually sound and thorough guide to approaching the issues of representativeness. The major problem with the issue of representativeness is not what criteria should be evaluated, but what remedies are available. In my experience, the majority of cases where probabilistic analysis is considered in environmental regulation/standard setting involve choosing a generic distribution to represent an essentially unknown population. That is, default distribution assumptions which can be employed in much the same way that standard point estimates are currently employed in (e.g.) the Superfund Program. Efforts such as the NHANES III project and other data gathering efforts on national and regional scales often provide data of excellent quality foe large scale populations, Notwithstanding that such data are often structured in a way which can permit information on specific subpopulations to be extracted in a representative fashion, we are rarely in a position to know who those subgroups should be in any given instance. While probabilistic analysis holds out the potential for realistic descriptions of the characteristics of real populations and their exposures, it has been, and, I believe, will continue to be rare for specific populations exposed at a given location to be characterized (other than possibly by their geographic location) in a way which will allow appropriate subpopulation data to be extracted from national/regional databases. If such populations were characterized and/or population-specific exposure data were collected in a focused study, then the issue of representativeness would become a more practical consideration. On the other hand, if such focused studies are not done, then there is little or no quantitative basis for considering whether national/regional population data are specific to the given population. Thus, in most cases the external data are likely to be "disjoint" with respect to the population of concern. In the absence of population-specific characterization (either with respect to demographics, or, preferably with respect to specific exposure), there does not appear to be an objective way of even identifying how the national/regional surrogate data may be biased with respect to the population of concern.

Having acknowledged this practical problem with deriving representative data exposure distributions, I am not sure that, from the standpoint of public health and risk-based regulation, it is necessarily wise that the population of concern be precisely characterized. The reason for this is that precise characterizations of populations are (as recognized in the checklists) precise with respect to individuals, and their location in space and time. Such information is only precise for a specific moment of time. Demographic and land use patterns change over time, and distributional data which are representative for a given population at a point in time may not be representative for the population at the same location several years or decades later. Risk-based regulatory decisions, on the other hand, are intended to be protective of the exposed population into the indefinite future. Too specific a description of a population of concern may, therefore, make a risk-based regulatory decision unprotective of future populations at the given location. Such considerations seem to argue for more generic tailoring of input exposure distributions to include an intentional component of true uncertainty to address the possible, but unknown values which might apply to future populations and land uses. It is not entirely clear how this should be addressed in quantitative terms, but as a starting point, it seems necessary for such generic descriptions to include the range of values which could reasonably be anticipated to apply to a generic population at a site. To the extent that such descriptions are biased with respect to the current population and/or land uses, that bias should (as appropriate) be toward including more of the high risk population than is already present at the site. For example, if the demographic make up of the potentially exposed population at a given site were such that there were few young children, the generic input distributions should assume that at some future time, the population could have a larger proportion of young children. It may not be necessary to assume that the national or regional demographics shift in a radical fashion (although over time such shifts, do, indeed, occur), but rather to assume that local demographic idiosyncracies are short-lived. Thus, if a specific locality or neighborhood is demographically skewed toward families with older children, or without children, it should be assumed that in the future, the demographics may shift such that the proportion of young children at the local level of a site reflects the overall state, or county proportion. Such assumptions should be based preferentially on analysis of regional population data, and, if such data are not available, on analysis of national data. One obvious problem with such an approach is that adjustments of current local demographics to current regional demographics to account for future local demographic shifts assumes that regional demographic patterns are more stable than local patterns. This may be true in general, but will not necessarily be true in any given instance.

Tiered Approach

The usual rationale for a tiered approach is that it saves the time and effort which would be needed to conduct population and/or site-specific analyses. Computational time per se, however, is not usually a limiting factor in such analyses. Site-specific data collection, on the other hand, is a major undertaking and is generally a limiting factor. Thus, if population-specific data are available and (as above) it is appropriate to base a risk-based regulatory decision on such data, there is no reason to employ a tiered approach to site-specific distributional descriptions. If, as above, regional-specific distributions are more appropriate for risk-based determinations, and such data are available, then, likewise, a tiered approach is not necessary. If, as is usually the case, population site, or region-specific data are not available, and national population-based data are available, such data may be appropriate as the basis for a screening approach. In considering the use of such data in a non-population-specific context, however, it must be asked to what extent the specific characteristics of the national data might be misleading for screening purposes. Specifically, are the details of the nation distribution in the extreme tails appropriate, even for screening purposes, for a given subpopulation? Given the screening nature of such an assessment, it may be more appropriate to generate and employ generic screening distributions which use quantitative approximations specifically intended for screening such as triangular, uniform and generalized distributions. Such distributions can also be applied when more complete national population distributions are not available. These distributions could describe, for example, relative minimum values, estimated 10% values, most likely values, estimated 90% values and relative maximum values. It is not necessarily clear that such generic distributions would not be more appropriate for screening purposes than national population-based data. Using such generic default screening distributions would have the additional advantage of establishing specific, and easily identified rebuttable presumptions which would form the starting point for site-specific modifications. Thus, starting from a default screening distribution, it might not be necessary to generate a complete site-specific distribution in order to move toward site/regional specificity. Rather, consideration of the default distribution may help focus the need for more specific information, and it might be realized that the most significant difference between the default assumption and the actual site/regional-specific distribution lies (e.g.) in the upper tail of the default distribution. Thus, it might be necessary only to collect data appropriate to modifying the 90% value in the default distribution.

APPENDIX G

POSTMEETING COMMENTS

David E. Burmaster

27 April 1998

Memorandum

To: Moderator, Participants, and Attendees --
Workshop on Selecting Input Distributions for Probabilistic Analyses

Via: Kate Schalk, ERG

From: David E. Burmaster

Subject: Thoughts and Comments After the Workshop in NYC

After much more reading and thinking, I remain staunchly opposed to letting the US EPA and its attorneys set a minimum value for any or all goodness-of-fit (GoF) tests such that an analyst may not use a fitted parametric distribution unless it achieves some minimum value for the GoF test.

In honesty, I must agree that GoF tests are useful in some circumstances, but they are not panaceas, they do have perverse properties, and they will slow or stop continued innovation in probabilistic risk assessment. The US EPA must NOT issue guidance, even though it is supposedly not binding, that sets a minimum value for a GoF statistic below which an analyst may not use a fitted parametric distribution in a simulation.

Here are my thoughts:

1. Re Data

For physiological data, many of the key data sets (e.g., height and weight) usually come from NHANES or related studies in which trained professionals use calibrated instruments to measure key variables (i.e., height and weight) in a clinic or a laboratory under standard conditions for a carefully chosen sample (i.e., adjusted for no shows) from a large population. These studies yield "blue-chip" data at a single point in time. These data, I believe, contain small but known measurement errors across the entire range of variability. At the extreme tails of the distributions for variability, the data do contain relatively large amounts of sampling error. Even with a sample of n = 1,000 people, any value above, say, the 95th percentile contains large amounts of sampling uncertainty. In general, the greater the percentile for variability and the smaller the sample size, the greater the (sampling) uncertainty in the extreme percentiles.

For behavioral and/or dietary data, many key data sets (e.g., drinking water ingestion, diet, and/or activity patterns) often come from 3-day studies in which the human subject recalls events during the previous days without the benefit of using calibrated instruments in a clinic or laboratory and not under standard conditions. Even though the researchers may have carefully selected a statistical sample from a large population, no one can know the accuracy or precision of the "measurements" reported by the subjects. These studies yield data of much less than "blue-chip" quality for a 3-day interval. These data, I believe, contain large and unknown measurement errors across the entire range of variability. At the extreme tails of the distributions for variability, the data also contain large amounts of sampling error. For a sample with n = 1,000, any value above, say, the 95th percentile contains large amounts of sampling uncertainty above and beyond the large amounts of measurement uncertainty. Again, the greater the percentile for variability and the smaller the sample size, the greater the (sampling) uncertainty in the extreme percentiles.

My conclusion from this? With all sample sizes, certainly with n < 1,000, I think the data are highly uncertain at high percentiles. I think it is inappropriate to eliminate a parametric model that captures the broad central range of the data (say, the central 90 percentiles of the data) just because a GoF test has a low result due to sampling error in the tails of the data. (This observation supports the idea that fitted parametric distributions may outperform EDFs at the tails of the data.) As Dale Hattis has written, use the process to inform the choice of parametric models -- not a mindless GoF test.

2. Re Fitted Parametric Distributions

As is well known:

a 6-parameter model will always fit data better than a 5-parameter model,
a 5-parameter model will always fit data better than a 4-parameter model,
a 4-parameter model will always fit data better than a 3-parameter model, and
a 3-parameter model will always fit data better than a 2-parameter model.

Thus, GoF tests always select models with more parameters than models with fewer parameters.

This perverse behavior contradicts Occam's Razor, a bedrock of quantitative science since the 13th century.

The venerable Method of Maximum Likelihood Estimation (MLE) offers an approach -- not the only approach -- to this problem. First, the analyst posits a set of nested models in which, for example, a n-parameter model is a special case of an (n+1)-parameter model -- and the (n+1)-parameter model is a special case of an (n+2)-parameter model. Using standard MLE techniques involving ratios of the likelihood functions for the nested models, the analyst can quantify whether the extra parameter(s) provide a sufficiently better fit to the data than does one of the simpler models to justify the computational complexity of the extra parameter(s).

3. Re Continued Innovation and Positive Incentives to Collect New Data and Develop New Methods

Over the last 15 years, the US EPA has issued innumerable "guidance" manuals that have had the perverse effect of stopping research and blocking innovation -- all in the name of "consistency."

In my opinion, our profession of risk assessment stands at a cross-road. The US EPA could specify, for example, all sorts of numeric criteria for GoF tests -- but the casualties would be (i) the continued development of new ideas and methods, especially the theory and practice of "second-order" parametric distributions and the theory and practice of "two-dimensional" simulations, and (ii) the use of expert elicitation and expert judgment.

I again urge the Agency print this Notice inside the front cover and inside the rear cover of each Issue Paper / Handbook / Guidance Manual, etc. related to probabilistic analyses -- and on the first Web page housing the electronic version of the Issue Paper / Handbook / Guidance Manual:

> This Issue Paper / Handbook / Guidance Manual contains guidelines and suggestions for use in probabilistic exposure assessments.
>
> Given the breadth and depth of probabilistic methods and statistics, and given the rapid development of new probabilistic methods, the Agency cannot list all the possible techniques that a risk assessor may use for a particular assessment.
>
> The US EPA emphatically encourages the development and application of new methods in exposure assessments and the collection of new data for exposure assessments, and nothing in this Issue Paper / Handbook / Guidance Manual can or should be construed as limiting the development or application of new methods and/or the collection of new data whose power and sophistication may rival, improve, or exceed the guidelines contained in this Issue Paper / Handbook / Guidance Manual.

References

Burmaster & Wilson, 1996

Burmaster, D.E. and A.M. Wilson, 1996, An Introduction to Second-Order Random Variables in Human Health Risk Assessment, Human and Ecological Risk Assessment, Volume 2, Number 4, pp 892 - 919

Burmaster & Thompson, 1997

Burmaster, D.E. and K.M. Thompson, 1997, Fitting Second-Order Parametric Distributions to Data Using Maximum Likelihood Estimation, Human and Ecological Risk Assessment, in press

P. Barry Ryan

Colleagues-

I read with interest the comments forwarded by Dr. David Burmaster regarding the conference from last week.

I would like to add a few similar words regarding the codification of any specific values for any specific goodness-of-fit (GOF) tests.

GOF tests, by their nature, are very restrictive in affording acceptance of a distribution. For example, the Kolmorgorov-Smirnoff test chooses the largest difference between the observed data and the theoretical ranking and tests using that. Unusual occurrences in data, minor contamination of by other distributions, etc., can cause rejection of distributions that otherwise pass the "duck test" (if it walks like a duck,...)even if one point looks a little more like a pigeon. The GOF test will end up rejecting pretty much everything leaving one with no choice but to use an EDF.

Unfortunately, EDFs are not readily amenable to analyses that lend a lot of insight (cf., Wallace, Duan, and Ziegenfus, 1994). If EPA codifies a fixed value, even in the guise of "guidance" pretty soon no pdf will be safe from legal wrangling.

We spent a long time at the workshop fussing over definitions of representativeness, sensitivity, etc., with little focus on the utility of the techniques. EPA may well be in the difficult position of having to defend everything from a legal perspective. However, the preoccupation with numbers often comes at the expense of insight. The role of probabilistic assessments is the latter. Our goal is to understand exposure and its influence on health, not to focus on a specific value of a GOF test statistic.

Somewhere in this document should be a statement equivalent to the one often seen in automobile commercials. "The material and techniques contained herein should only be used by professionals familiar with the nuances of the problem at hand and the techniques used, their limitations, and strengths." I object to the cookbook approach to this type of assessments.

I will now step down off my soapbox.

P. Barry Ryan
Professor, Exposure Assessment and Environmental Chemistry
Rollins School of Public Health
Emory University
1518 Clifton Road NE
Atlanta, Georgia 30322
(404) 727-3826 (Voice)
(404) 727-8744 (Fax-Work)
bryan@sph.emory.edu

APPENDIX H

PRESENTATION MATERIALS

EPA
United States
Environmental Protection
Agency

RAF Workshop on Selecting Input Distributions for Probabilistic Risk Assessment

RISK ASSESSMENT FORUM

Forum Background

- Formed from Recommendations in "Risk Assessment in the Federal Government: Managing the Process" -NAS (1983)
- Members Include Senior Scientists from EPA Laboratories, Regions, and Program Offices
- Products Include Guidance Documents and Technical Reports on Difficult Risk Assessment Issues

Forces Affecting EPA's Use of Probabilistic Risk Assessment

- EPA Policy and Guidelines
- SAB and NAS Recommendations
- Risk-Based Legislation (e.g., Food Quality Protection Act of 1996)

Past Forum Involvement

- 1993 *ad hoc* Technical Panel on Monte Carlo Analysis
- 1996 Workshop on Monte Carlo Analysis
- 1997 Policy for Use of Probabilistic Analysis in Risk Assessment at the U. S. Environmental Protection Agency

EPA United States Environmental Protection Agency

RISK ASSESSMENT FORUM

Policy for Use of Probabilistic Analysis in Risk Assessment

- Policy Issued May 15, 1997
  - “...Such Probabilistic Analysis Techniques as Monte Carlo Analysis...can be Viable Statistical Tools for Analyzing Variability and Uncertainty in Risk Assessment”
- Eight Conditions for Acceptance
- Policy Implementation Activities

Conditions for Acceptance

5. Information for each input and output distribution is to be provided in the report. This includes tabular and graphical representations of the distributions (e.g., PDF and CDF plots) that indicate the location of any point estimates of interest (e.g., mean, median, 95th percentile). *The selection of distributions is to be explained and justified.* For both the input and output distributions, variability and uncertainty are to be differentiated where possible.

Implementation Activities

- Immediate Activities
  - “Summary Report for Workshop on Monte Carlo Analysis” (EPA/630/R-96/010)
  - “Guiding Principles for Monte Carlo Analysis (EPA/630/R-97/001)

- Follow-Up Activities
  - *Distributions for Selected Exposure Factors*
  - Agency Training on Probabilistic Analysis Methods
  - Detailed Technical Guidance on Quantitative Uncertainty Analysis

Implementation Activities Cont'd

- Longer Term Activities
  - Development or Revision of Guidelines and Models
  - Update of Exposure Factors Handbook
  - Modeling Support Group
  - Dose-Response Applications of Probabilistic Assessment

EPA United States Environmental Protection Agency

RISK ASSESSMENT FORUM

Selecting Input Distributions to Represent Exposure Factors

- Fitting Distributions to Data That Are Not Perfectly Representative of the Scenario Under Study
- Using the Empirical Distribution Function or Resampling Techniques Instead of Parametric Distributions

WORKSHOP ON SELECTING INPUT DISTRIBUTIONS FOR PROBABILISTIC ASSESSMENT

H. Christopher Frey, Ph.D.
Workshop Chair and Participant

Department of Civil Engineering
North Carolina State University
Raleigh, NC 27695

Presented At:
U.S. Environmental Protection Agency
New York City

April 21-22, 1998

FOCUS OF WORKSHOP

- Produce a written document
- Provide advice
- Representativeness
- Empirical versus Parametric Distributions

Workshop Approach

- Plenary discussions
- "Brainwriting" activity
- Writing assignments: individuals and/or small groups

NC STATE

Brainwriting Activity

- Small groups (4-6)
- Trigger question
- Silent activity
- Interactive
- Simultaneously prepare written responses (pen and paper!)
- Exchange pages
- React to previous comments and add more
- Continue until exhausted and/or time is up

ROLE OF OBSERVERS

- The proceedings of this workshop are public information
- This workshop deals with inherently technical subject matter, and it is not intended as a short course or tutorial
- The live panel discussion may be highly technical at times
- Designated times on both days for observer comments and questions
- 30 minutes total each day
- May impose time limits on individual speakers

Who is the Audience?

- As experts, we tend to talk to each other in jargon.
- Potential difficulties communicating across disciplines
- Try to present concepts in plain language if possible.
- This will be especially important in the written record.
- Audience for the written record is “the assessor”

EPA's Expectations

- "Representativeness" is a higher priority than EDF vs. PDF, but both should be addressed
- Focus on the issue papers. The RTI report is for background only and is not the subject of this workshop.
- Focus on technical issues
- Keep in mind the audience for the framework document that will result from this effort

NC STATE

Representativeness:
What should we focus on?

- **The Issue Paper? Any significant comments?**
- **Sensitivity: How important is potential lack of representativeness? How do we evaluate this?**
- **Adjustments: Are there reaonable ways to adjust or extrapolate from the sample to the population of interest?**

Representativeness: How do you make adjustments?

- Population Characteristics
- Spatial Characteristics
- Temporal Characteristics

- What methods are available now?
- What activities are needed in the short term to develop or explore new methods?
- What longer term research is needed to develop new methods?

NC STATE

Other Issues?

- What, if anything, is not covered in the issue paper or the charge that is on topic and should be addressed?

Other Issues?

- **Some Examples Based Upon Pre-Meeting Comments:**
  - Role of expert judgment and Bayesian methods, especially in making adjustments.
  - Representativeness of Probabilistic Model Outputs (is exposure or risk estimate representative even if each individual input is representative?)
  - Role of representativeness in a default or generic assessment
  - Role of measurement process needs more attention

General Questions

- What do we know today that can be applied to answering the questions or providing guidance?
- What short term studies (e.g., numerical experiments) could be conducted to answer the question or provide additional guidance?
- What longer term research may be needed to answer the question or provide additional guidance?

NC STATE

REPRESENTATIVENESS OF EXPOSURE FACTORS DATA

Jacqueline Moya

National Center for Environmental Assessment

Workshop on Selecting Input Distributions for Probabilistic Analysis

New York

April 21 - 22, 1998

Outline

- Purpose
- Framework for Inferences
- Components of Representativeness
- Checklists
- Attempting to Improve Representativeness

Purpose

- Discuss the concepts
- Discuss how to evaluate representativeness
- Discuss how to address representativeness

General Framework for Inferences

- **Representativeness:**
  Comfort with which one can draw inferences

- **Population:**
  Set of units defined by individuals' characteristics

Components of Representativeness

Define Population of Concern in Terms of:

- Individual Characteristics
- Spatial Characteristics
- Temporal Characteristics

Components of Representativeness

- Assessor's Population of Concern

  population for which assessment is being conducted

- Surrogate Population

  population used by the assessor when data on the population of concern are not available

- Population of Concern for the Surrogate Study

  population for which the surrogate study was designed

- Population Sampled

  a sample from the population of concern of the surrogate study

Components of Representativeness

- **External Components:**

  Population of Concern vs Surrogate Population

- **Internal Components:**

  Population Sampled vs Population of Concern for the Surrogate Study

Components of Representativeness

External to Study:

How well does the surrogate population represent the population of concern?

Components of Representativeness

Internal to Study:

- How well do individuals sampled represent population of concern for the study?
- How well do actual number of respondents represent the sampled population?
- How well does the measured value represent the true value for the measured unit?

Checklists

- Assessing representativeness of population sampled vs population of concern for the surrogate study
- Assessing representativeness of surrogate population vs assessor's population of concern (individual, spatial, and temporal characteristics)

Attempting to Improve Representativeness

- Adjustments to account for differences in population characteristics
- Adjustments to account for time-unit differences

Issues

- Are the checklists useful tools?
- How does one evaluate the importance of the differences between two populations?
- How should extrapolations be made?

Discussion Topic 2

Empirical Distribution Functions vs Analytic Distribution Functions

Assessing Quality of Fit for Analytic Distributions

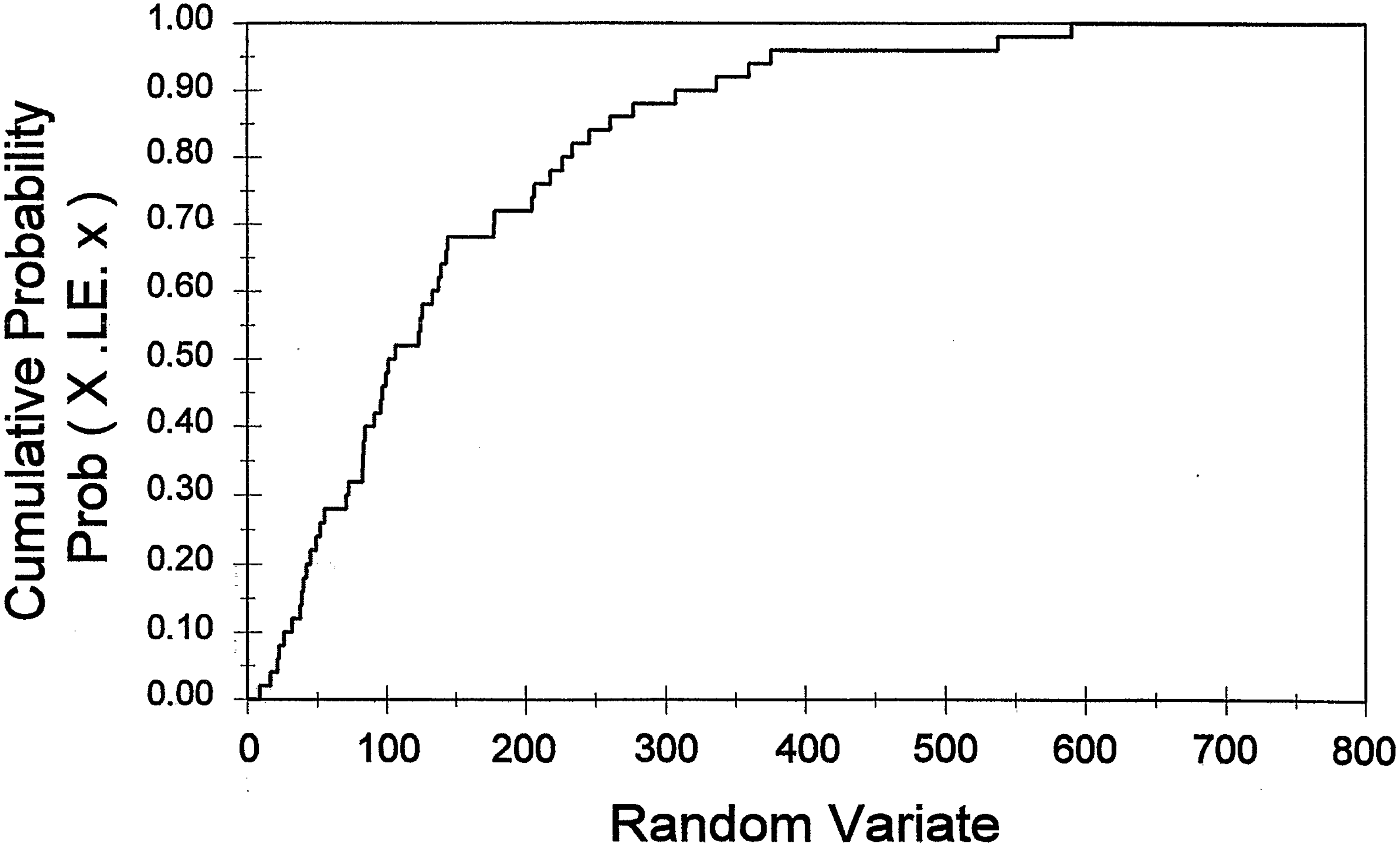
Empirical Distribution Function (EDF)

Given representative data, X = $\{x_1, x_2, \cdots, x_n\}$, sorted from smallest to largest, $x_1 \le x_2 \le \cdots \le x_n$, the EDF is the cumulative distribution function defined by

$$\hat{F}(x) = \frac{number\ \ of\ x_k \le x}{n} \qquad or \qquad \hat{F}(x) = \frac{1}{n}\sum_{k=1}^{n} H(x - x_k)$$

where H(u) is the unit step function which jumps from 0 to 1 when u ≥ 0. The values of the EDF are the discrete set of cumulative probabilities $(0, 1/n, 2/n, \cdots, n/n)$.

Figure 1. Example of EDF



Properties of the EDF

- Values between any two consecutive samples, x_k and x_{k+1} cannot be simulated, nor can values smaller than the sample minimum, x_1, or larger than the sample maximum, x_n, be generated, i.e., $x \geq x_1$ and $x \leq x_n$

- The mean of the EDF is equal to the sample mean. The variance of the EDF mean is always smaller than the variance of the sample mean; it is equal to $(n-1)/n$ times the variance of the sample mean.

- The variance of the EDF is equal to $(n-1)/n$ times the sample variance.

- Expected values of simulated EDF percentiles are equal to the sample percentiles.

- If the underlying distribution is skewed to the right, the EDF will tend to under-estimate the true mean and variance.

Variations of the EDF

Linearized EDF

linearly extrapolating between observations

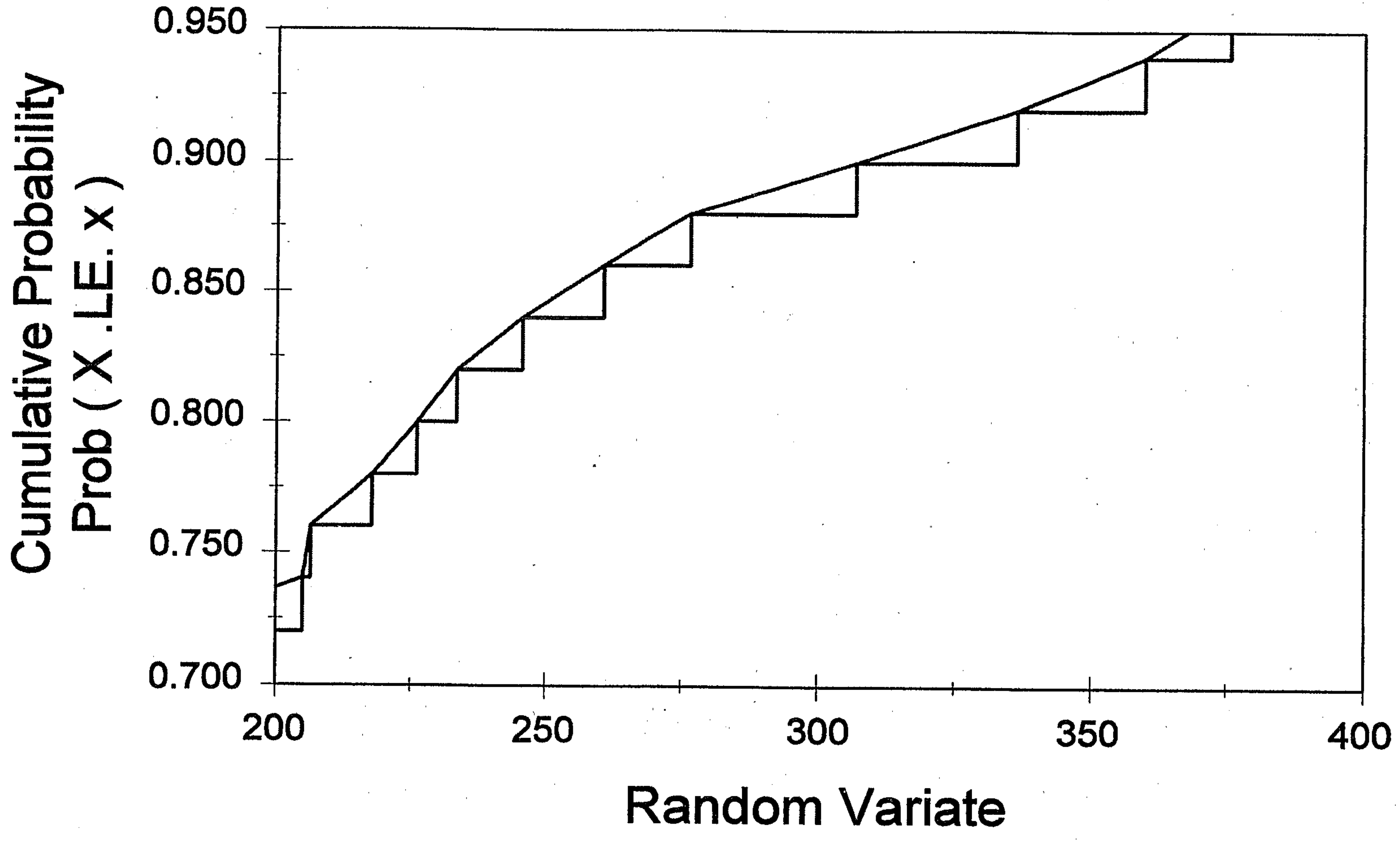
Extended EDF

based on expert judgement, adding lower & upper tails to the data to reflect “**a more realistic range**” of the exposure variable (EDFs produce tails that are too short)

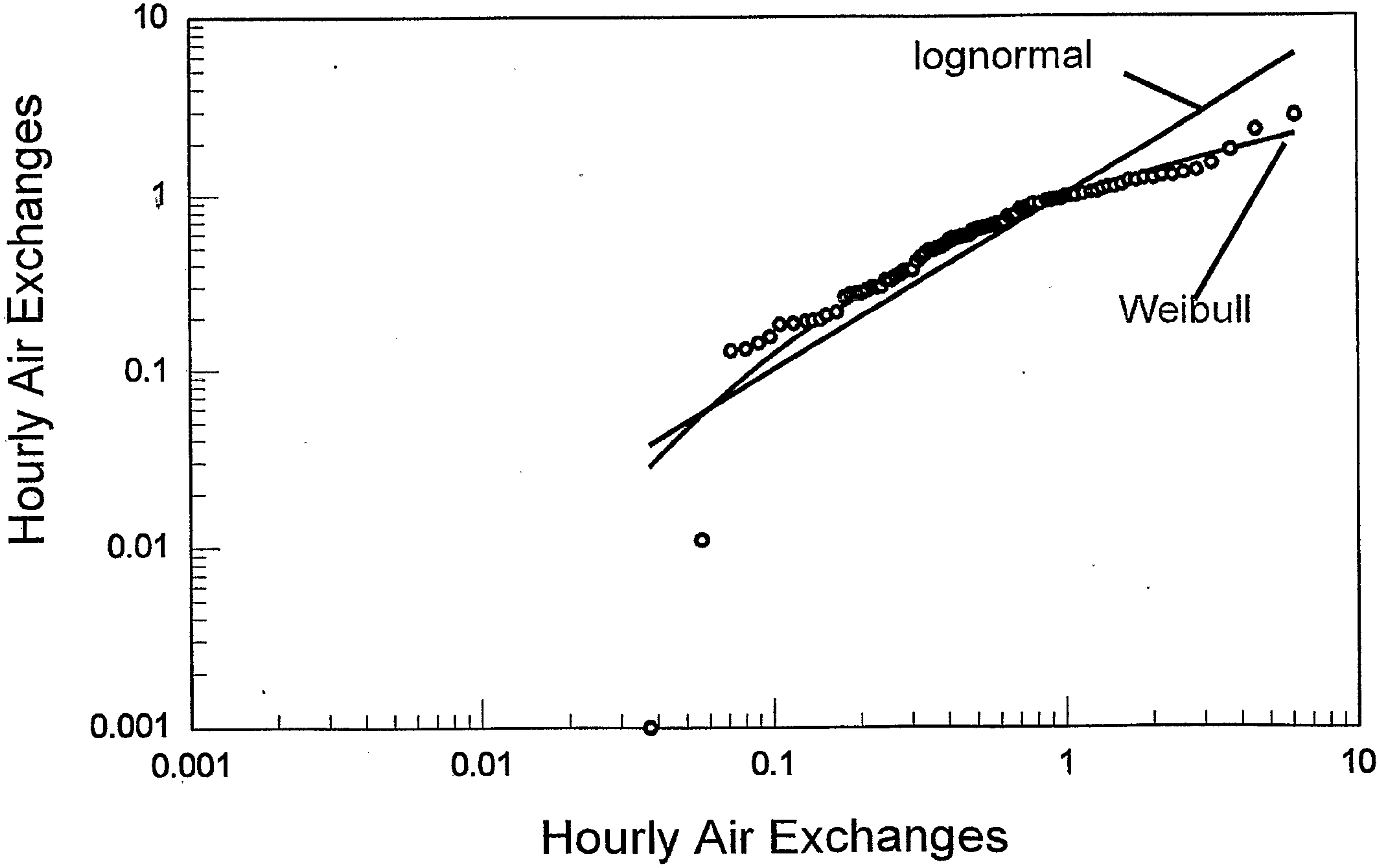
Mixed Exponential

based on extreme value theory, adding an exponential upper tail to the EDF to model the exponential behavior of many continuous, unbounded distributions

Figure 4. Comparison of Basic EDF and Linearly Interpolated EDF



Comparison of Fitted Lognormal and Weibull Distributions to ACH Data



Statistic	ACH Sample N = 90	EDF	Linearized EDF	Best Fit Weibull PDF
mean	0.6822	0.6821	0.6747	0.6782
variance	0.2387	0.2358	0.2089	0.2479
skewness	1.4638	1.4890	1.2426	1.2329
kurtosis	6.6290	6.7845	5.6966	4.9668
5%	0.1334	0.1320	0.1307	0.0881
10%	0.1839	0.1840	0.1840	0.1452
50%	0.6020	0.6160	0.6032	0.5691
90%	1.2423	1.2390	1.2398	1.3592
95%	1.3556	1.3820	1.3600	1.6450

EDF Questions

- Are there circumstances in which an EDF is preferred over a TDF?

- Are there situations in which an EDF should not be used?

- When an EDF is used, should the linearized, extended or mixed versions be used?

Goodness of Fit Questions

Generally, we should pick the simplest analytic distribution not rejected by the data...... **But,** rejection dependents on the statistic chosen and an arbitrary level of statistical significance.

- What role should the GoF statistic and its p-value (when it is available) play in that judgment?

- What role should graphical assessments of fit play?

- When none of the standard distributions fit well, should we investigate more flexible families of distributions, e.g. four parameter gamma, four parameter F, mixtures, etc.?

BAYESIAN ANALYSIS OF VARIABILITY AND UNCERTAINTY OF ARSENIC CONCENTRATIONS IN U.S. PUBLIC WATER SUPPLIES

John R. Lockwood and Mark J. Schervish
Department of Statistics

Patrick L. Gurian
Department of Engineering & Public Poiicy

Mitchell J. Small
Departments of Engineering & Public Policy and
Civil & Environmental Eng.

Carnegie Mellon University

presented at
7th Annual Meeting of ISEA
Session 1: Approaches to Uncertainty Analysis
Research Triangle Park, NC
November 3, 1997

Sponsor:
U.S. EPA OGWDW
(Usual disclaimer . . .)

OBJECTIVES

- Illustrate use of *Bayesian statistical methods*
  - variability in an exposure factor (arsenic concentration) is represented by a probability distribution model
  - uncertainty is characterized by the probability distribution function of the model parameters
- Illustrate use of probability distribution model with *covariates (explanatory variables)*
  - allowing extrapolation to different target populations

Methodology

- Probability model with covariates
  - lognormal distribution with mean of ln's a function of
    - region,
    - source type (sw vs. gw), and
    - size of utility (population served)
  - constant variance of ln's
- Bayesian methodology
  - prior distribution for model parameters
  - posterior distribution computed using Markov Chain Monte Carlo (MCMC)
    - necessitated by model complexity and BDL data

Arsenic Occurrence Databases

Database	Sample Locations	Source Type	Number of Sites	Detection Limit, µg/L	Percentage Below Detection Limit
WITAF National Arsenic Occurrence Survey (NAOS)	Raw water	Surface and groundwater	441	0.5	37%
EPA National Inorganics and Radionuclides Survey (NIRS)	Finished water	Groundwater	982	5	93%
Association of California Water Agencies	Raw water and finished water	Surface and groundwater	1,542	1	37%
12 State Data Reported to EPA	Raw water and finished water	Surface and groundwater	>11,000	Varies	65%

MODEL SPECIFICATION

$$Y_{ij} = \mu_i + \beta x_{ij} + \gamma g_{ij} + \epsilon_{ij}$$

- Y_{ij} is the natural logarithm of arsenic concentration in μg/L at j^{th} source in i^{th} region

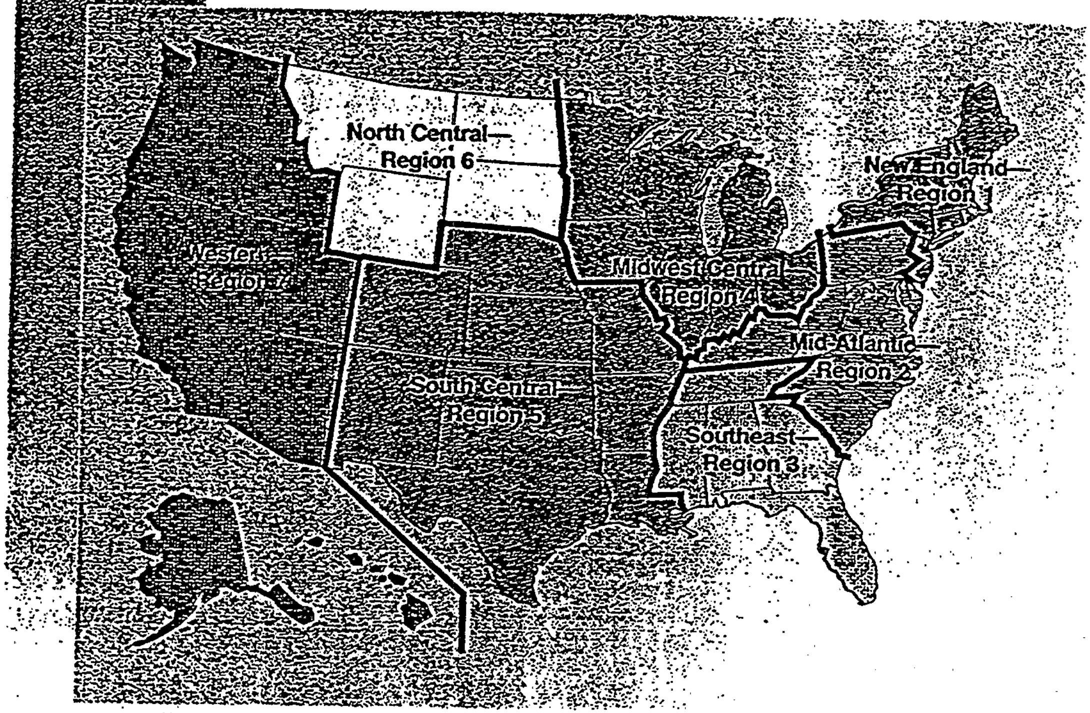
- μ_i is a constant for i^{th} region, where i ranges over the seven geographical regions specified in NAOS

- x_{ij} is the natural logarithm of the population served by j^{th} source in i^{th} region (an indicator of the size and flow rate of the utility source)

- g_{ij} is 0 if j^{th} source in i^{th} region is a surface water source and 1 if it is a ground water source

- ϵ_{ij} represents those sources of random variation present at the j^{th} source in i^{th} region but not captured by the covariates in the model.

FIGURE 2 US geographic regions based on arsenic NOFs



Source: Frey and Edwards, 1997

DISTRIBUTIONAL ASSUMPTIONS

In the model

$$Y_{ij} = \mu_i + \beta x_{ij} + \gamma g_{ij} + \epsilon_{ij}$$

it is assumed that

$$\epsilon_{ij} \sim N(0, \sigma_\epsilon^2) \qquad \forall i, j$$

$$\mu_i \sim N(\psi, \tau^2) \qquad i = 1, \ldots, 7$$

That is, μ_i are sampled from a parent normal distribution (hierarchical model).

The normality assumption of ϵ_{ij} implies that conditional on all parameters,

$$Y_{ij} \sim N(\mu_i + \beta x_{ij} + \gamma g_{ij}, \sigma_\epsilon^2)$$

BAYESIAN METHODOLOGY

- Probability model:

$$X \sim f_{X|\Theta}(x|\theta)$$

where Θ are parameters $\Theta \sim f_{\Theta}(\theta)$

- Begin with prior distribution $f_{\Theta}^{0}(\theta)$
- Observe sample $X = \vec{x}_s$
- Compute posterior distribution

$$f_{\Theta|X}(\theta|\vec{x}_s) = \frac{f_{X|\Theta}(\vec{x}_s|\theta)f_{\Theta}^{0}(\theta)}{\int_{\Theta} f_{X|\Theta}(\vec{x}_s|\theta)f_{\Theta}^{0}(\theta)d\theta}$$

PRIOR DISTRIBUTIONS

Without substantive prior knowledge about parameters of hierarchical model, our priors were diffuse:

$$\begin{aligned} \psi &\sim N(0,3^2) \\ \beta &\sim N(0,10^2) \\ \gamma &\sim N(0,10^2) \\ log(\sigma^2) &\sim N(0,10^2) \\ log(\tau^2) &\sim N(0,10^2) \end{aligned}$$

These parameters are assumed independent *a priori*, but are dependent in the posterior.

POSTERIOR ESTIMATES

Parameter	P.M.	P.S.D.
μ_1	-3.13	0.65
μ_2	-3.50	0.61
μ_3	-3.62	0.61
μ_4	-1.76	0.57
μ_5	-1.84	0.59
μ_6	-1.04	0.66
μ_7	-1.41	0.62
σ^2	2.23	0.21
ψ	-2.27	0.74
τ^2	1.76	1.76
β	0.21	0.05
γ	0.14	0.19

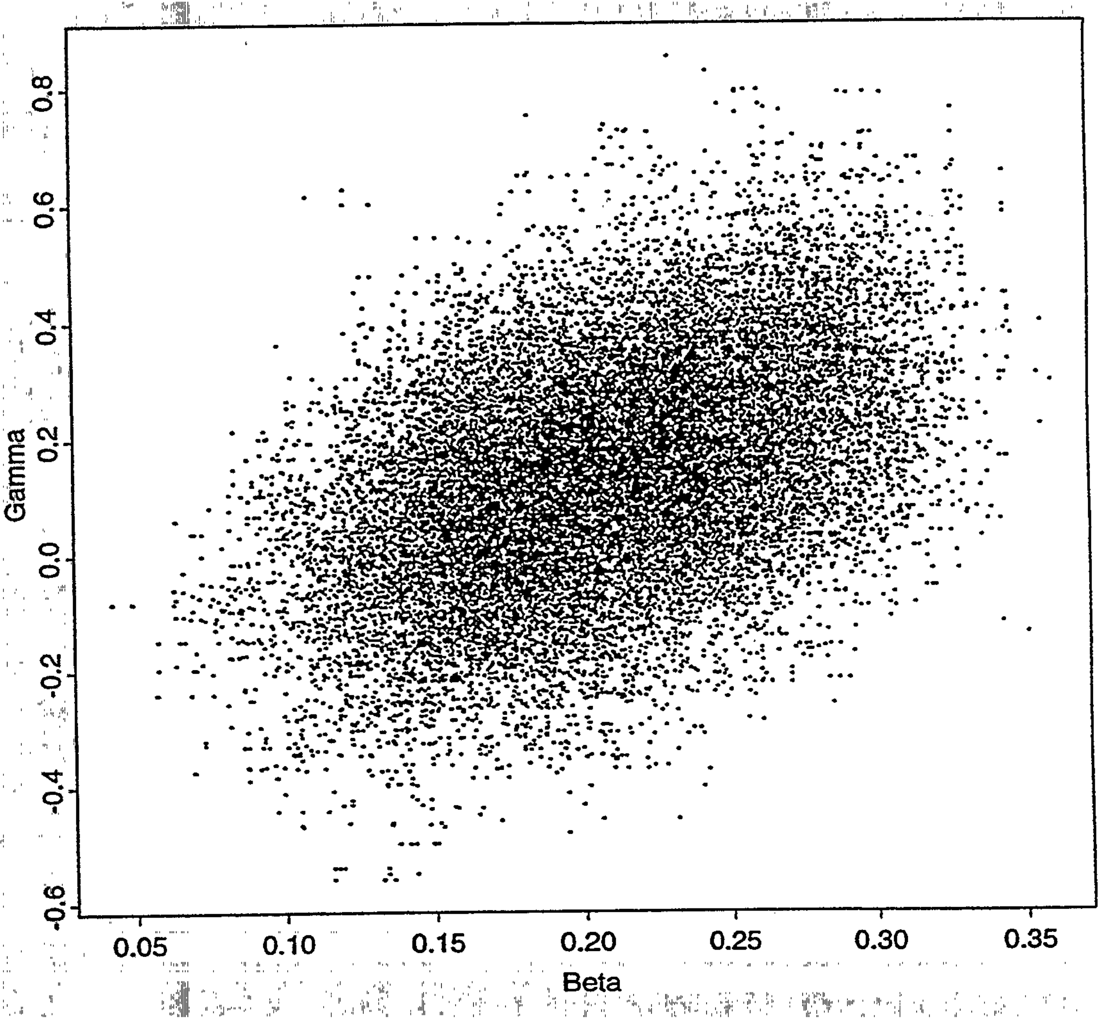


Figure 2: Scatterplot of γ versus β from a sample of size 30000 from the joint posterior distribution.

NATIONAL DISTRIBUTION & UNCERTAINTY

The national distribution of arsenic concentration measurements is the mixture of all the distributions from the individual sites:

$$F_{\text{National}} = \frac{1}{N} \sum_{\text{All sites } i} F_i,$$

where N is the total number of sites in the nation.

Similarly for our estimates:

$$\hat{F}_{\text{National}} = \sum_{\text{All sampled sites } i} w_i \hat{F}_i,$$

where w_i is a weight indicating how much of the nation is represented by site i.

However, $\hat{F}_i$ is uncertain due to uncertainty in model parameters. The posterior uncertainty in $\hat{F}_i$ is characterized by the many (equally likely) $\hat{F}_{i,j}$ obtained by evaluating $\hat{F}_i$ with the parameters in MCMC sample j.

We can then compute the mean, cdf, median, 5th percentile, 95th percentile, etc. of the distribution of $\hat{F}_i$

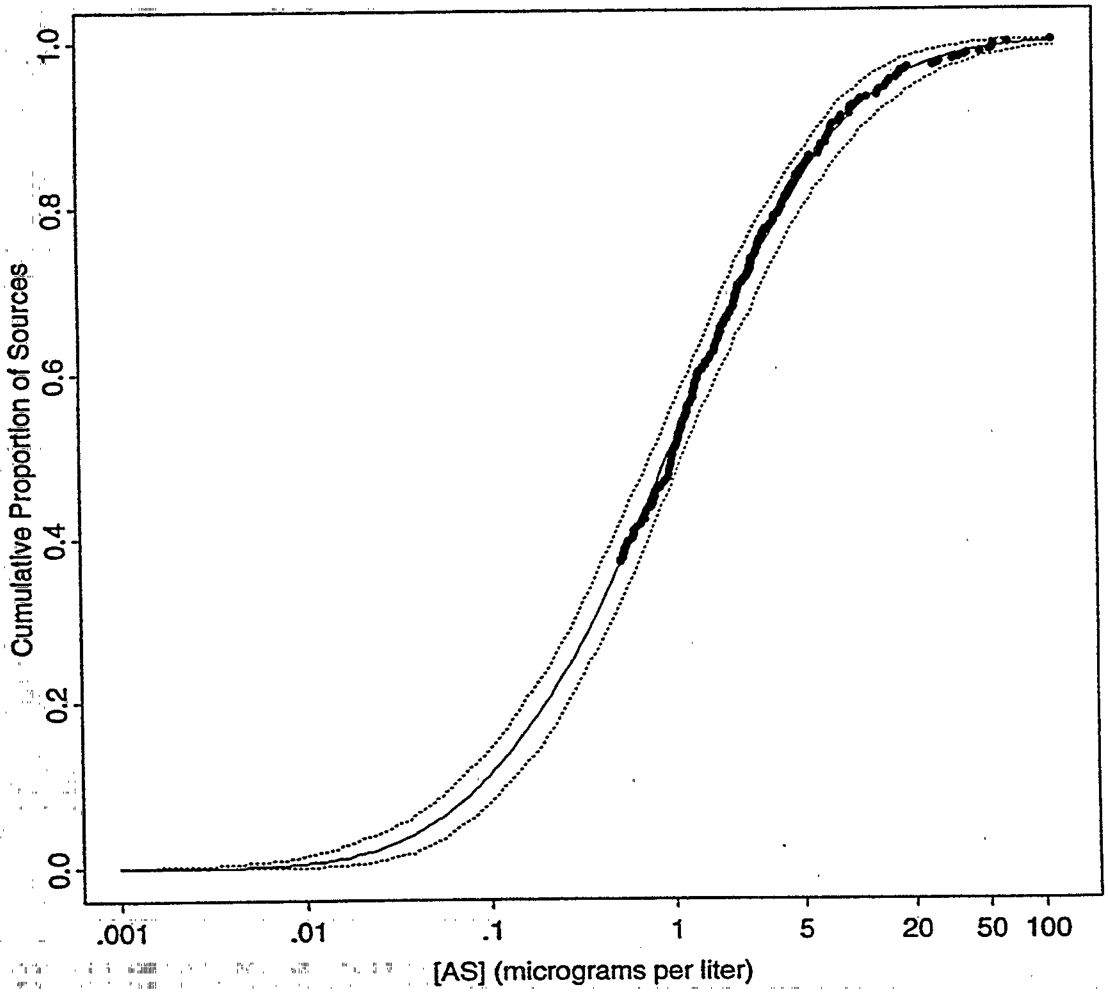


Figure 3: Posterior cumulative distribution function of national arsenic occurrence in source water with 90% credible bounds and uncensored NAOS data overlayed.

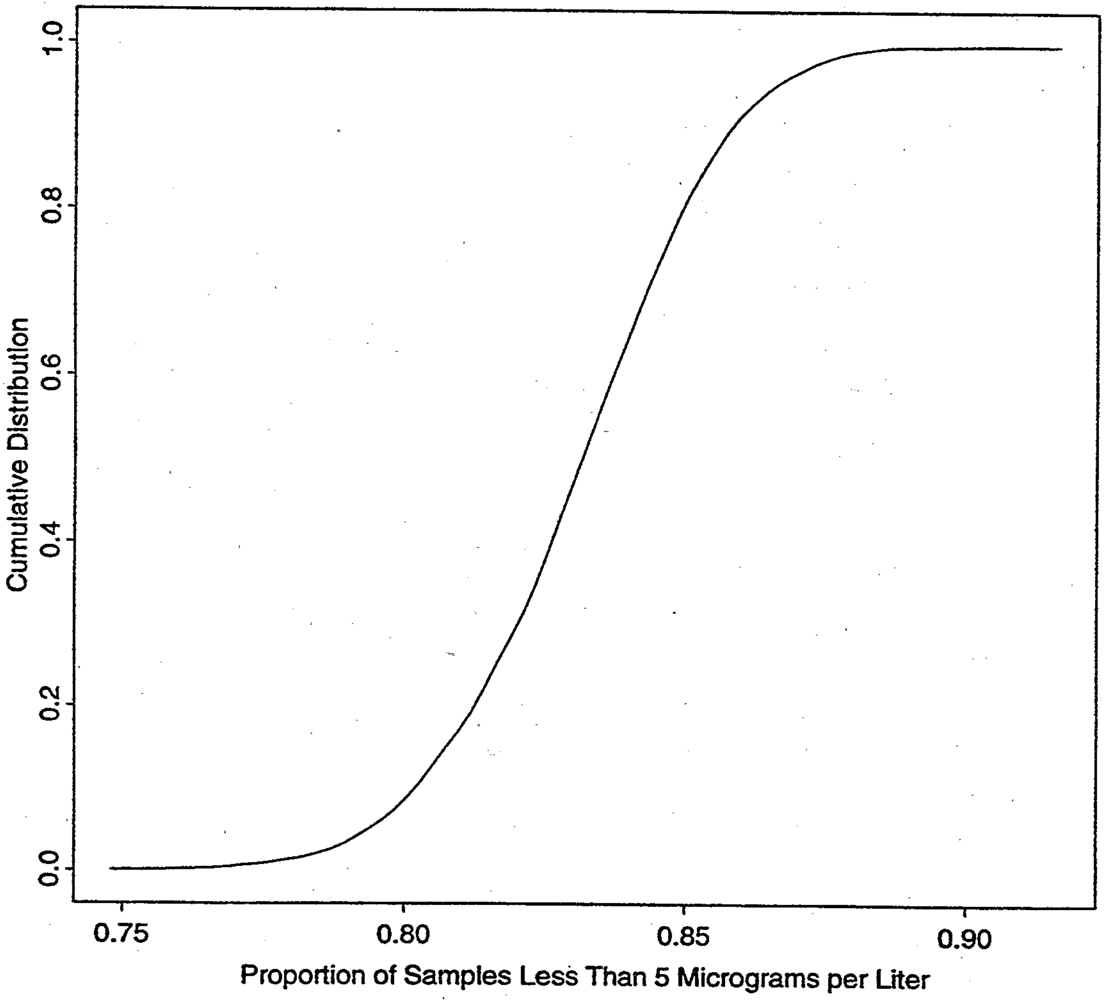


Figure 4: Posterior cumulative distribution function of the proportion of national arsenic occurrence less than 5 μg/L.

POSTERIOR ESTIMATES: Alternative Model

Supposing the γ should be positive, we kept all priors the same as before except took the prior for $log(\gamma)$ to be $N(0, 10^2)$.

Parameter	P.M.	P.S.D.
μ_1	-2.78	0.55
μ_2	-3.17	0.51
μ_3	-3.27	0.49
μ_4	-1.42	0.44
μ_5	-1.50	0.48
μ_6	-0.71	0.54
μ_7	-1.04	0.48
σ^2	2.22	0.20
ψ	-1.94	0.65
τ^2	1.74	1.75
β	0.18	0.04
γ	0.03	0.07

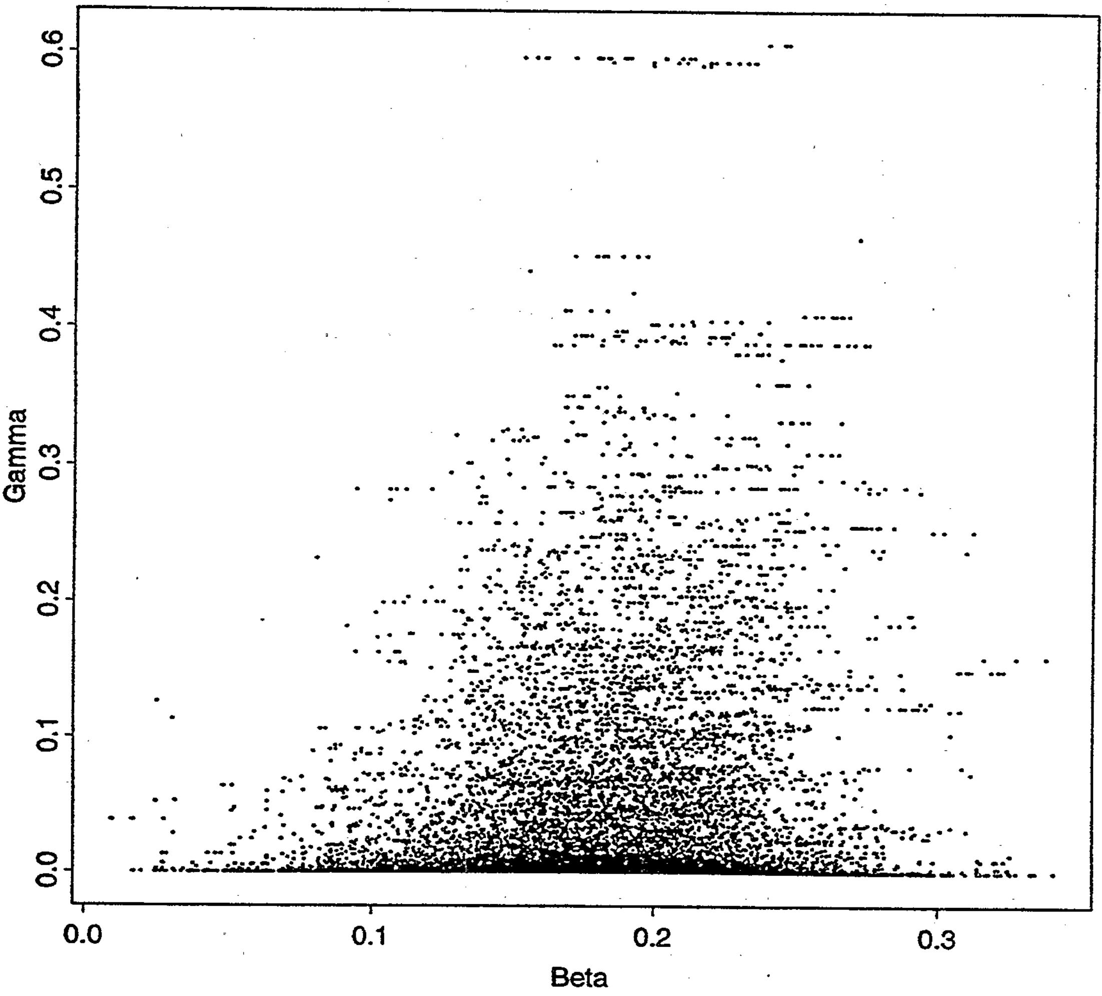


Figure 7: Scatterplot of γ versus β from a sample of size 30000 from the joint posterior distribution when γ is forced to be postive.

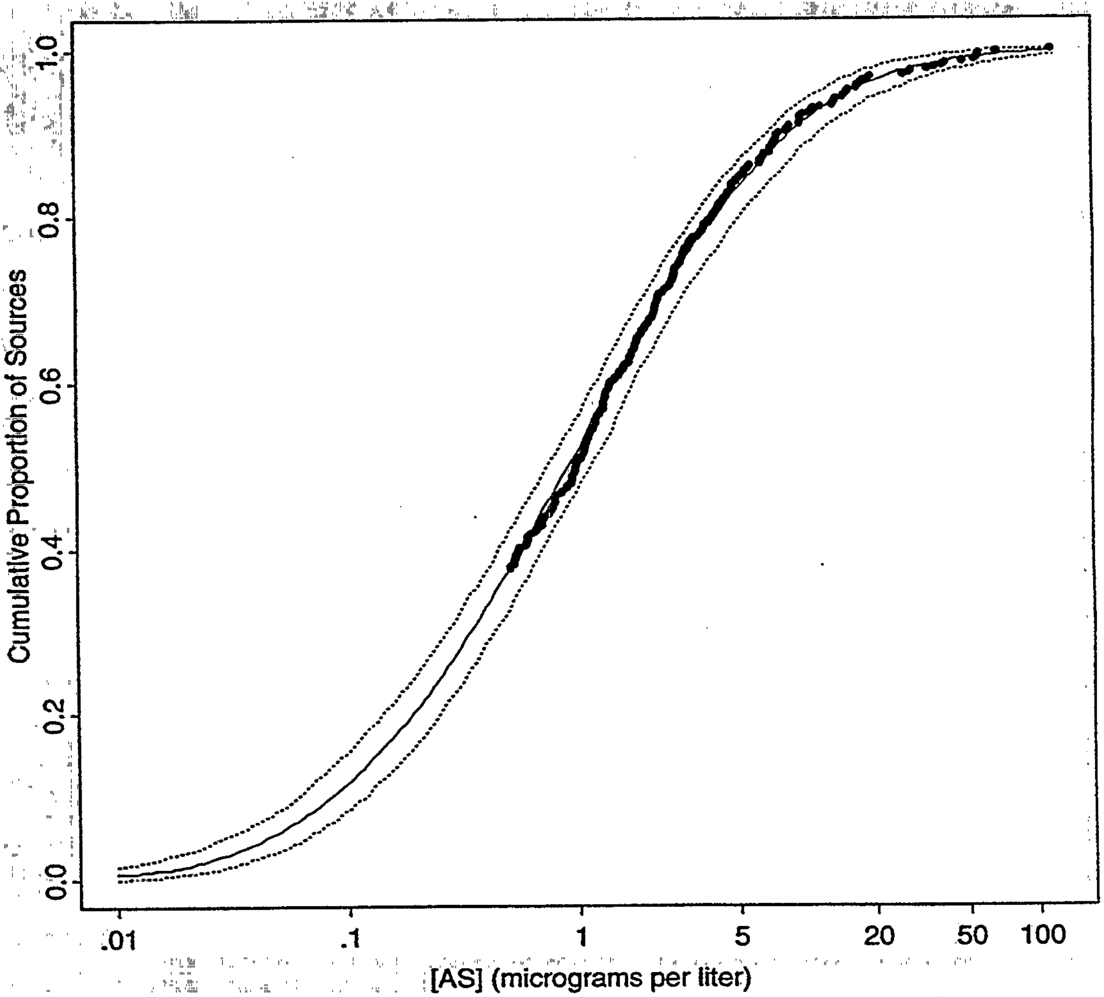


Figure 8: Posterior cumulative distribution function of national arsenic occurrence in source water with 90% credible bounds and uncensored NAOS data overlayed. Plot based on posterior when γ is forced to be positive.

SUMMARY

- Bayesian methodology provides a powerful method for characterizing variability and uncertainty in exposure factors
  - effect of alternative priors can be investigated in a diagnostic manner
  - though don't try this at home alone *(without a competent statistician)*
- Probability distribution model with covariates provides insights, and a basis for extrapolation to other targeted populations or subpopulations.

Bayesian Analysis of Variability and Uncertainty of Arsenic Concentrations in U.S. Public Water Supplies

John R. Lockwood
Mark J. Schervish
Department of Statistics

Patrick L. Gurian
Department of Engineering & Public Policy

Mitchell J. Small
Departments of Engineering & Public Policy
and Civil & Environmental Engineering

Carnegie Mellon University

presented at

Seventh Annual Meeting of the International Society of Exposure Analysis
Research Triangle Park, NC
November 3, 1997

(Session 1. Approaches to Uncertainty Analysis)

Bayesian Analysis of Variability and Uncertainty of Arsenic Concentrations in U.S. Public Water Supplies

John R. Lockwood[1],
Mark J. Schervish[1],
Patrick L. Gurian[2]
and Mitchell J. Small[3]

The risk of skin and other possible cancers associated with arsenic in drinking water has made this problem a top priority for research and regulation for the U.S. EPA, as part of implementation of the Safe Drinking Water Act amendments of 1986 and 1996. To assess the costs, benefits and residual risks of alternative maximum contaminant levels (MCL's) for arsenic, it is important to characterize the current national distribution of arsenic concentrations in the U.S. water supply. This paper describes a Bayesian methodology for estimating this distribution and its dependence on covariates, including the source region, type (surface vs. ground water) and size of the source. The uncertainty of the fitted distribution is also described, thereby depicting the uncertainty in the proportion of utilities with concentrations above a given MCL. This paper describes the first stage of this assessment, based on a sample of concentrations from source water drawn by utilities. Subsequent analyses will incorporate the distribution and effectiveness of current treatment practices for reducing arsenic, and include available data sets of finished water quality to estimate the arsenic concentration distribution in water supplied to consumers.

Using arsenic concentration data for source (raw) water reported by 441 utilities from the National Arsenic Occurrence Survey (NAOS) (Frey and Edwards, 1997), we fit a Bayesian model to describe arsenic concentrations based on source characteristics. The model allows for both the formation of a national estimate of arsenic occurrence and the quantification of the uncertainty associated with this estimate. The specification of the model is

$$Y_{ij} = \mu_i + \beta x_{ij} + \gamma g_{ij} + \epsilon_{ij}$$

where

- Y_{ij} is the natural logarithm of arsenic concentration in μg/L at j^{th} source in i^{th} region
- μ_i is a constant for i^{th} region, where i ranges over the seven geographical regions specified in NAOS
- x_{ij} is the natural logarithm of the population served by j^{th} source in i^{th} region (an indicator of the size and flow rate of the utility source)
- g_{ij} is 0 if j^{th} source in i^{th} region is a surface water source and 1 if it is a ground water source

[1] Department of Statistics, Carnegie Mellon University.

[2] Department of Engineering and Public Policy, Carnegie Mellon University.

[3] Departments of Engineering and Public Policy and Civil and Environmental Engineering, Carnegie Mellon University.

- ϵ_{ij} represents those sources of random variation present at the j^{th} source in i^{th} region but not captured by the covariates in the model.

Furthermore, we model the values μ_i as independent normal random variables with mean ψ and variance τ^2. The national distribution of arsenic in source water is thus modeled as a mixture of lognormals with the mean of the log-concentration equal to $\mu_i + \beta x_{ij} + \gamma g_{ij}$ and the standard deviation of the log-concentration equal to σ. The resulting distribution depends upon the number of utilities in each of the seven regions (i), their service populations x and the respective numbers drawing water from surface ($g_{ij} = 0$) vs. ground ($g_{ij} = 1$) water (for now, the sample is assumed to be representative of the national distribution, though the predicted distribution can be readily modified to reflect a different distribution of the covariates in the target population).

To characterize the uncertainty of the fitted national distribution, we use vague prior distributions for the parameters ψ, τ, β, γ, σ and employ the Markov Chain Monte Carlo methodology (Gilks et al., 1996) to compute and simulate realizations from the posterior distribution of the parameters. Posterior uncertainty distributions of all quantities of interest can be calculated from these realizations.

Table 1 lists the posterior means and posterior standard deviations for the fitted model parameters. The mean values indicate that

- arsenic concentrations are generally higher in the west than in the east (the posterior means of μ_4, μ_5, μ_6 and μ_7 are greater than the posterior means of μ_1, μ_2 and μ_3)

- arsenic concentrations tend to be higher in source waters of larger utilities (the posterior mean of β is positive)

- arsenic concentrations are higher in ground water than in surface water (the posterior mean of γ is positive, though there is significant uncertainty in this result since the posterior standard deviation of γ is greater than the posterior mean)

The uncertainty in the fitted national distribution is characterized by the standard deviations of the parameters shown in Table 1 and by the covariance of the parameters in the posterior joint distribution. Figures 1 and 2 illustrate this covariance for two of the parameter pairs: (β, ψ) and (β, γ), respectively. These covariances are of the type that commonly arise in parameter estimation; for example, the positive association between higher β (which results in higher predicted arsenic concentrations) and lower ψ (which corresponds to lower values of the μ_i and lower predicted arsenic concentrations) is necessary to maintain the match to the observed sample values.

The national distribution is synthesized by sampling the joint parameter space (i.e, the points in Figures 1 and 2 and the associated points for the other model parameters) to generate many possible distributions. For each, the cumulative distribution function (cdf) at a particular value of the arsenic concentration ($\exp(Y)$) is computed as the average of the predicted cdf's for each measurement in the original sample of 441, based on its model covariates (or, the covariates for each utility in the target population, if these differ from the sample). The multiple cdf's generated from the parameter space describe the uncertainty of the national variability distribution. The median of the uncertainty distribution is one

Table 1: Posterior means and standard deviations of parameters. The regions (subscripts) are 1=New England, 2=Mid-Atlantic, 3=Southeast, 4=Midwest Central, 5=South Central, 6=North Central, 7=West.

Parameter	Posterior Mean	Posterior Standard Deviation
μ_1	-3.18	0.67
μ_2	-3.51	0.62
μ_3	-3.66	0.63
μ_4	-1.78	0.59
μ_5	-1.89	0.62
μ_6	-1.10	0.67
μ_7	-1.47	0.64
σ^2	2.17	0.20
ψ	-2.30	0.76
τ^2	1.74	1.77
β	0.21	0.05
γ	0.14	0.19

choice for a single estimate of the national distribution. This median distribution is shown in Figure 3, along with corresponding 5th and 95th percentiles and the observed distribution of the original data set. The fitted distribution closely matches the observed distribution, including the result that 37% of the sample is at or below the arsenic detection limit of 0.5 μg/L. The full uncertainty distribution for the proportion of the national population below one particular value of the arsenic concentration (5 μg/L) is shown in Figure 4, where this proportion is indicated to range from about 0.79 – 0.87, with a median of 0.83. This characterizes the uncertainty in the proportion of utilities requiring treatment of their source water to meet an MCL of 5 μg/L.

Acknowledgment: This work is sponsored by the U.S. EPA Office of Ground Water and Drinking Water, Standards and Risk Management Division. The paper has not been subject to EPA peer review, and the views expressed are solely those of the authors.

References

Frey, M. M. and M. A. Edwards (1997). Survey arsenic occurrence. *Jour. AWWA*, **89**(3), 105-117.

Gilks, W. R., S. Richardson and D. J. Spiegelhalter, eds (1996). *Markov Chain Monte Carlo in Practice.* Chapman and Hall, London.

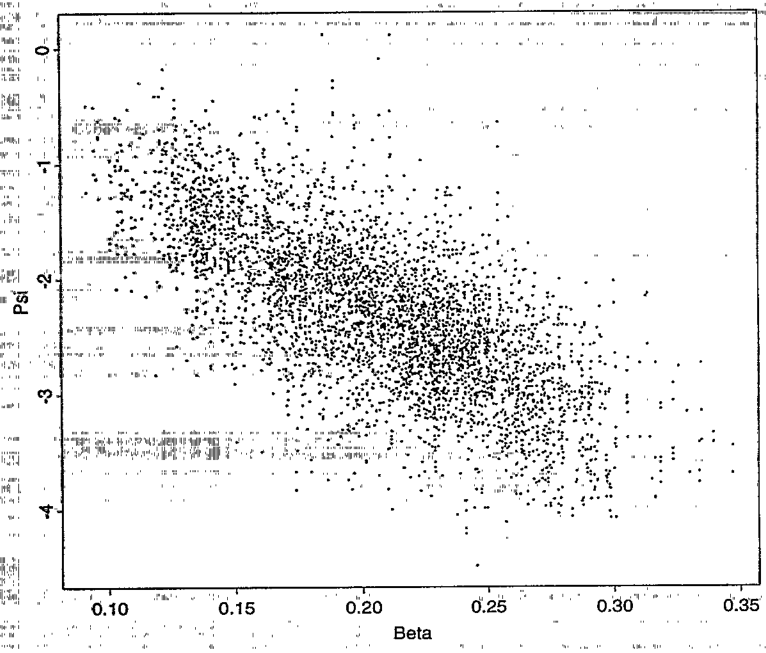


Figure 1: Scatterplot of ψ versus β from a sample of size 5000 from the joint posterior distribution

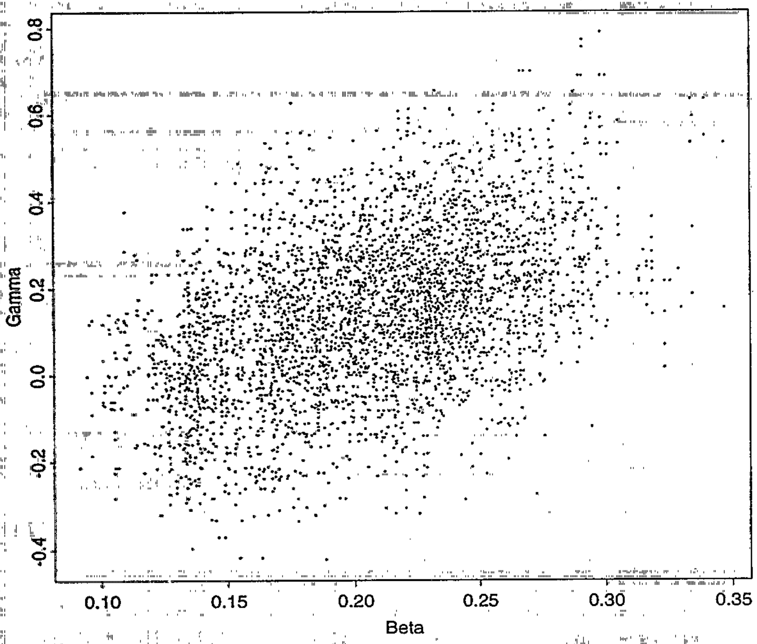


Figure 2: Scatterplot of γ versus β from a sample of size 5000 from the joint posterior distribution

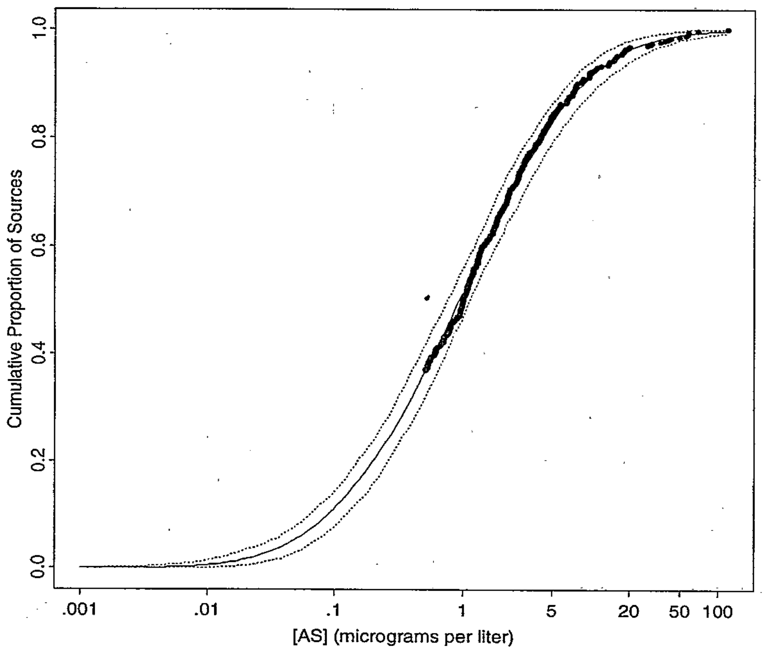


Figure 3: Posterior cumulative distribution function of national arsenic occurrence in source water with 90% credible bounds and uncensored NAOS data overlayed.

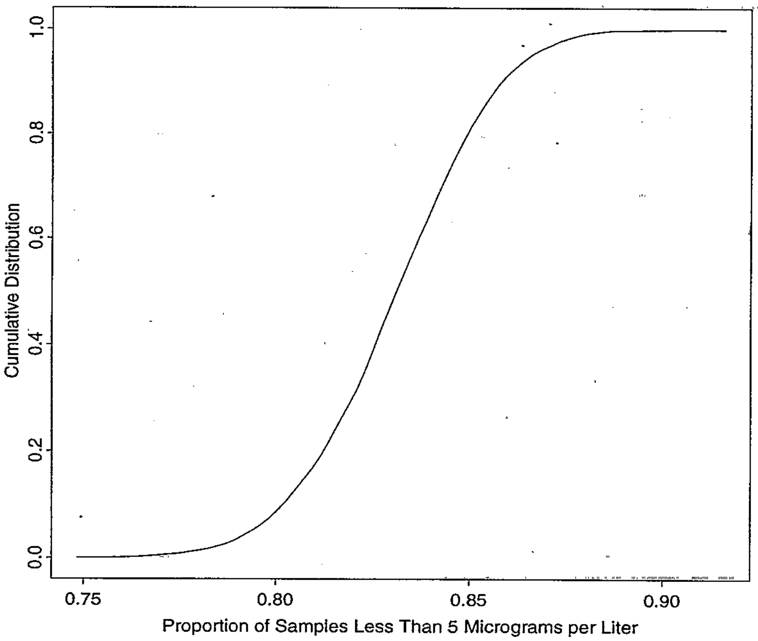


Figure 4: Posterior cumulative distribution function of the proportion of national arsenic occurrence less than 5 μg/L

☆U.S. GOVERNMENT PRINTING OFFICE: 1999 - 750-101/00039