6

Discussion and Recommendations

This section discusses applicability on a large scale of the methodology presented here to other factors in the Exposure Factors Handbook (EFH). Data quality issues are discussed first, then recommendations are presented.

6.1 Adequacy of Data

As defined in Section 1, a statistical methodology is a combination of an experimental design or data set, a class of models, and an approach to inference. Although each of these three factors is important, their relative importance as determinants of the overall quality of the output is the same as the order given. That is, the quality of the data is obviously the most important factor, the quality of the models is second in importance, and the approach to inference is third (Cox, 1990; Johnson, 1978).

The greatest possible gains in overall risk assessment quality would come from designing and conducting a survey of the population of interest for each risk assessment. Ideally, individuals selected from the population by probability-based sampling would be monitored for periods of time, which would allow both long-term and short-term parameter estimates. Duplicate diet techniques would be employed, whereby exact copies of all foods and beverages consumed would be obtained, weighed, catalogued, and chemically analyzed for each subject. Similarly representative direct tactics would be employed for other routes of exposure for the same sample subjects. Probability-based surveys are uniquely qualified to produce representative data on the population of interest and entail minimal assumptions.

However, the customized survey approach will rarely be used. In most cases, exposure assessors must do the best they can, working with data from diverse sources, summarized in the fashion of the EFH. For some of the EFH factors, it may be possible to update the key studies.

In some cases, EFH data are extremely limited and may consist of only a single number, such as an estimated mean. In such a case, one is tempted to claim that the data are inadequate for choosing distributions. However, the risk assessor may not have this luxury. In many cases, something must be done, no matter how limited the data are, or even in a complete absence of data. In fact, cases of such limited data are precisely the kind where quantification of uncertainty is most important. Expert judgment may have to be substituted for data, and sweeping, apparently unwarranted, assumptions may be necessary. Sensitivity analysis is almost essential in such a case. In implementing the sensitivity analyses, two or three plausible assignments should be made for the distribution of the factor, and a corresponding number of risk assessment simulations should be done, based on each assignment. Of course, if F factors each require D different distributions for sensitivity analysis, then F*D separate risk assessment simulations are required, which could be prohibitively expensive.

Given only a mean and standard deviation, or only a mean and 99th percentile, one would produce the corresponding gamma, lognormal, and Weibull distributions. Since each would fit the two given numbers perfectly, one would have no data-based method for preferring one of the three, and each would have to be used in risk assessment simulations to investigate sensitivity of conclusions to the type of distribution. If only a point estimate, such as a mean, were available, one would try to obtain a plausible population coefficient of variation (CV) or standard deviation by considering similar factors or eliciting expert judgment. Then, one would determine the gamma, lognormal, and Weibull distributions with the given mean and standard deviation and recommend that all three be used in risk assessment simulations to investigate sensitivity of conclusions.

To conclude the discussion of data adequacy, it is important to acknowledge again that situations will arise where distributional assumptions are difficult to justify. In some of these cases, empirical distributions may be used.

6.2 Application of Methodology to Other Exposure Factors

The remainder of this section concerns the applicability of this methodology to other exposure factors of the EFH. The available EFH data summaries can be roughly classified into four cases: (1) six or more percentiles available, (2) three to five statistics available, (3) two statistics available, and (4) at most one statistic available. Raw data are rarely available. However, if raw data were available, it would probably be treated as case 1, unless the sample sizes were very small.

6.2.1 Case 1: Percentile Data

Summary of Methodology

- Models: 12-model hierarchy based on generalized F with point mass at zero

- Estimation: maximum likelihood

- Goodness-of-fit (GOF) tests: chi-square and likelihood ratio tests (LRTs)

- Uncertainty: asymptotic normality for large samples, bootstrap or normalized likelihood for small samples

Many EFH data summaries contain six or more empirical percentiles for the given population and factor. In many cases, other information also is provided, which may include a sample mean, standard deviation, sample size, and percent exposed or percent consuming. This is referred to as the percentile case, even though other information besides percentiles is also usually available.

Because more percentiles than moments are available, it seems reasonable to focus the analysis on the percentiles, using the moment information as a check or validation on the distribution estimated from the percentiles. However, the possibility of tailoring the inference to all the available information is not ruled out. For example, the tap water data of Section 3 include nine empirical percentiles, the sample mean, and sample standard deviation for each age group. Model parameters could be estimated to minimize the average percent error in all 11 of these quantities. The resulting nonstandard estimate would not have a nice textbook distribution, but simulation or bootstrap techniques could be used to approximate its distribution to obtain GOF tests and uncertainty parameter distributions.

The joint asymptotic distribution of any specified sample percentiles is known to be multivariate normal, with known means, variances, and covariances (Serfling, 1980). The joint asymptotic distribution of specified sample moments is also known (Serfling, 1980). Conceivably, the joint asymptotic distribution of specified percentiles and moments also could be determined. This would make it possible to apply a conventional type of asymptotic analysis that takes into account all of the available sample percentile and moment information.

If six or more percentiles are available, the methods applied in Section 3 to the tap water data are recommended. Specifically, use maximum likelihood estimation (MLE) to fit the five-parameter generalized F distribution and all of the special cases identified in Sections 1 and 2 and used in Section 3. For formal GOF, use both the chi-square test of absolute fit and the LRT of fit relative to the five-parameter model. To obtain distributions for parameter uncertainty, use asymptotic normality for large samples, and use bootstrapping or the normalized likelihood for small samples. Ideally, simulation studies would be used to at least check on coverage probabilities associated with the uncertainty analysis.

6.2.2 Case 2: Three to Five Statistics Available

Summary of Methodology

Models: two-parameter gamma, lognormal, and Weibull

Estimation: minimize average absolute percent error in the available statistics

GOF tests: bootstrapping

Uncertainty: bootstrapping

Because the available information is quite limited, consideration should be given to obtaining the raw data.

If only three to five statistics are available, information is very limited, and it seems important to use all available quantities in the estimation process. Such limited data also make it difficult to justify going beyond the two-parameter models. Accordingly, fitting the two-parameter gamma, lognormal, and Weibull models, using estimation to minimize the average absolute percent error in all available quantities, is recommended. (With four or five statistics available, it would also be possible to fit the generalized gamma, in addition to the two-parameter models.)

If the original sample size n is known, then bootstrapping can be used to obtain *p*-values for GOF as well as to obtain parameter uncertainty distributions. To illustrate these applications of bootstrapping, assume that three statistics are originally available: for example, the mean, standard deviation, and 90th percentile. Parameters have been estimated for each of the three models (gamma, lognormal, and Weibull) by minimizing the average absolute percent error. To apply bootstrapping, first generate 1,000 random samples of size n from the estimated (gamma, lognormal, or Weibull) distribution. For each sample, calculate the mean, standard deviation, and 90th percentile. Also for each sample, determine the minimized average absolute percent error (MAAPE) and note which parameter values achieve the minimum. Rank these 1,000 MAAPEs from largest to smallest. The *p*-value for GOF is determined by the location of the original MAAPE among the 1,000 ordered simulated MAAPEs. For instance, if the original MAAPE is between the 47th and 48th largest ordered simulated MAAPEs, then the *p*-value for GOF is 0.048. The parameter uncertainty distribution for each model is simply the discrete distribution that places mass 0.001 on each of the simulated parameter pairs for that model. The possibility of bias in the bootstrapped parameter pairs should be checked. If necessary, such bias can be removed by a simple translation so that the mean of the parameter uncertainty distribution is equal to the original estimated parameter vector.

This bootstrap approach can be used for each of the three types of models. GOF *p*-values can be used to decide whether model uncertainty requires that more than one of the three types of models be used for risk assessment.

6.2.3 Case 3: Two Statistics Available

Summary of Methodology

- Models: two-parameter gamma, lognormal, and Weibull
- Estimation: exact agreement with the available statistics
- GOF tests: not applicable
- Uncertainty: bootstrap the available statistics for each model

Another fairly common EFH situation involves only two summary statistics, such as a mean and upper percentile, or a mean and standard deviation. We will assume for illustrative purposes that the mean and standard deviation are available. If bio-physico-chemical considerations do not dictate the type of model, then determining the two-parameter gamma, lognormal, and Weibull distributions that agree with the given information is recommended. Because of the considerable model uncertainty, at least the first two types of models should be used in risk assessment. In some cases, such as CV<50%, as in

Section 5 for inhalation rates, the differences between the models may be negligible relative to the overall risk assessment so that use of any one of the models may be sufficient.

Because of data limitations, the models fit the available data perfectly and formal GOF tests are not possible.

For parameter uncertainty distributions for each type of model, bootstrapping from the estimated model can be used to obtain a distribution of parameter uncertainty, as described in Section 6.2.2. That is, using the original estimated model parameters, 1,000 random samples of the original size are generated and summarized in terms of the same two quantities, mean and standard deviation. For each such simulated pair, the model agreeing with the mean and standard deviation is determined. This yields parameter uncertainty distributions.

6.2.4 Case 4: At Most, One Statistic Available

If this situation arises, it will have to be treated on a case-by-case basis, as described in the fifth paragraph of Section 6.1. Subjective, even Bayesian methods, would seem to be required, using expert judgment and analogies with other similar factors to hypothesize models and parameter distributions.

6.2.5 Topics for Future Research

In Section 1.1, we discuss briefly two important problems related to the iid (identically and independently distributed) assumption: modeling data from complex survey designs and the need to account for correlations among exposure factors. While both issues were beyond the scope of the present study, their importance cannot be overstated. Since risk assessors often lack raw data and must work with published data summaries that may not be properly weighted, it would be useful to investigate (perhaps by simulation) the magnitudes and nature of inaccuracies that arise by ignoring various aspects of sample designs. Further, it would be interesting to examine whether these biases might be differentially reflected in different PDF models and/or estimation procedures; in particular, it would be useful to compare the robustness of the nonparametric density estimators to the parametric probability density functions (PDF) models.

Because many exposures, especially through dietary intake, are strongly correlated, multivariate PDF modeling may be preferable to the univariate approach presented here. While multivariate models are more realistic, their complexity makes them much more difficult to fit, estimate, and validate, and they require considerably more data than their univariate counterparts. Nonetheless, efforts should be made to extend the topics covered in this report to the multivariate case. Recent availability of user-friendly software for implementing multivariate parametric PDFs in Monte Carlo risk assessment models (Millard, 1998; Millard and Neerchal, 1999) suggests that if the data are available and the limitations and requirements properly understood, multivariate PDF models could be utilized by risk assessors who have a basic understanding of statistical methods.

Finally, it should be noted that this report does not address temporal correlations within individuals. Frequently, risk assessors will want to model, longitudinally, an individual's exposure to one or more risk factors from birth to some advanced age. However, it is likely that assessors will have to utilize cross-sectional exposure data reported for discrete age classes. While the methods described in this report can be used to fit parametric PDFs to such data, there is an implicit assumption that the age-specific exposure distributions are mutually independent. In reality, a person's quantile values in the various age-specific distributions will be correlated. Thus, a person who is in the first quartile of meat ingestion in the j^{th} age class is more likely to be in the first quartile of the $j+1^{th}$ age class than is a person who was in the third quartile of meat ingestion in the j^{th} age class. This problem is similar to the multivariate exposure factor issue, just discussed, and should have a similar solution. It is important to investigate and solve both in a manner that allows risk assessors to develop more realistic and flexible models.