A System for Fitting Distributions to Exposure Factor Data

The system has components for models, estimation, assessment of fit, and uncertainty. In a production process to analyze a large number of Exposure Factors Handbook (EFH) datasets, a reduced set of options may be appropriate. Appendix B illustrates some pertinent calculations using tap water consumption data for adults over age 65.

2.1 Models

The system is based on a 16-model hierarchy whose most general model is a five-parameter generalized F distribution with a point mass at zero. The point mass at zero represents the proportion of the population that is not exposed or does not consume. A smaller set of models might be used to analyze a large number of EFH datasets. The first 8 models of our 16-model hierarchy (numbers of adjustable parameters) are:

# Generalized F (4)
# Generalized gamma (3)
# Burr (3)
# Gamma, lognormal, Weibull, log-logistic (2)
# Exponential (1)

These models are discussed in Chapter 2 of Kalbfleisch and Prentice (1980). The generalized F and generalized gamma models are power-transformed central F (Pearson type VI) and gamma random variables. The degrees of freedom parameters for the F distribution do not have to be integers but can be any positive numbers. Several two-parameter models are specified because two-parameter models are most commonly used.

The other eight models are obtained from those above by incorporating a point mass at zero. This increases the number of adjustable parameters by one. The point mass is simply the probability that a randomly selected population member is not exposed. Two additional models that may occasionally be useful are the normal distribution for approximately symmetric data and the beta distribution for bounded populations with known bounds.

For a process to be applied to a large number of factors from the EFH, the use of the basic two-parameter gamma, lognormal, and Weibull distributions is desirable for simplicity. In some cases, it may be necessary to use a more general model to achieve satisfactory fit to the data. For instance, these three models are unified within the three-parameter generalized gamma family, which includes them as special cases. The need for a more general model might occur with large datasets, datasets exhibiting multiple peaks or modes, or exposure factors where some individuals are not exposed. A large dataset might require a model with more than two parameters to achieve adequate fit. A mixture of two-parameter models may be needed in a multimodal situation. The inclusion of a parameter representing a point mass at zero exposure may be required to fit a population containing a nonnegligible proportion of unexposed individuals. Occasionally, a dataset may defy fit by standard textbook parametric models, and recourse to the empirical distribution may be appropriate.

2.2 Methods of Estimation

# Maximum likelihood estimation (MLE)
# Minimum chi-square (MCS) estimation
# Weighted least squares (WLS) estimation
# Minimum distance estimation (MDE)
# Method of moments (MOM) estimation
# Meta-analysis
# Regression on age and other covariates

These methods of estimation are defined in Kendall and Stuart (1979) and Kotz and Johnson (1985); maximum likelihood was discussed in some detail in Section 1.2.

In classifying methods of statistical estimation, it is useful to take an operations research or optimization point of view. The statistician or modeler summarizes the objectives of estimation as a real-valued criterion function. For the first four cases above, the criteria are the likelihood function, a chi-square measure, a weighted sum of squares of errors, and a distance function. Having formulated the problem in terms of a criterion function, the modeler proceeds to estimate parameters to optimize (maximize or minimize) the criterion function. This typically leads to a calculus problem, that is, the problem of finding a critical point where the partial derivatives of the criterion function with respect to the parameters are equal to zero. Unfortunately, in most cases of interest, one cannot just write down the partial derivatives and find their roots using simple algebra. A trial and error method is usually required, using an iterative search routine starting from an approximate solution to the problem.

Optimization is a major branch of applied mathematics. Obtaining and validating solutions to multidimensional optimization problems is not simple. Good overviews of optimization are given by Chambers (1973) and Press et al. (1992).

2.2.1 Maximum Likelihood Estimation

MLE is applicable to raw data and to percentile data. A likelihood for the data is obtained using the probability model in conjunction with assumptions regarding independence or dependence. The MLE is the parameter vector that maximizes the likelihood. Loosely speaking, the MLE is the parameter vector for which the data at hand are most likely. The MLE is the most plausible value for the parameter, if plausibility is measured by the likelihood.

2.2.2 Minimum Chi-Square Estimation

To use MCS estimation, it is necessary to group the data into categories. The categories can be defined by selected percentiles, so that MCS is applicable to percentile data as well as raw data. A certain number of the data points fall into each category. These are called the observed counts and are denoted by the symbol O. Under the model assumptions, for a given set of parameter values, a corresponding expected (E) number of sample points fall into each category. The chi-square value is the summation over the categories of $(O-E)^2/E$. In some cases, an O is used in the denominator instead of

E. This is referred to as the modified chi-square statistic. In either case, the MCS estimate is the parameter vector that minimizes the chi-square value.

2.2.3 Weighted Least Squares and Minimum Distance Estimation

WLS, or regression, estimates are chosen to minimize a weighted sum of squared discrepancies between model and data. Usually the weights are inversely proportional to (estimated) variances. WLS estimators include several MDEs as special cases and are applicable to either raw data or percentile data.

2.2.4 Method of Moments Estimation

The MOM produces estimates of parameters so as to achieve exact or approximate agreement with specified sample moments. Hence, the criterion function is some measure of distance between model-based and empirical moments. For example, the MOM can be applied by estimating the parameters of a two-parameter model to provide exact agreement with the sample mean and standard deviation. Generally speaking, the MOM is less efficient than the other methods mentioned above and is not widely used. However, if the only available information is a sample mean and standard deviation, there are few other options. The reader is referred to Kendall and Stuart (1979) for more detailed information about this method.

2.2.5 Estimation by Meta-Analysis

Meta-analysis is a set of techniques to synthesize information from multiple studies. For instance, suppose there are estimated means and standard deviations for the same or similar populations from multiple studies. It is possible to use analysis of variance techniques to estimate an overall mean, as well as between-study, within-study, and total variation. The MOM can then be used to determine gamma, lognormal, and Weibull distributions with mean and variance equal to the estimated overall mean and total variance. This technique is used in Section 5 to estimate a distribution for long-term inhalation rates.

2.2.6 Regression on Age and Other Covariates

Parametric regression methods similar to those used in the field of clinical biostatistics provide a promising technique to unify and summarize environmental exposure distributions across age groups. This might entail some additional compromise of fit at the level of the individual age group, but the resulting simplicity and unity of summaries may be worth the price. Risk assessment simulations may also be simplified by programming a formula for repeated use in different age groups. The approach works best if more general models (e.g., at least the generalized gamma) are used as the default. For example, in the case of population mobility (discussed in Section 4), all three of the two-parameter models (gamma, lognormal, and Weibull) are needed to obtain best fit to the data from the different age groups. Thus, the regression approach would be simplified, in this case, by using the generalized gamma, which contains all three distributions as special cases.

2.2.7 Distributions of Related Test Statistics

Associated with each type of estimation is additional machinery needed to approximate the probability distribution of the statistics obtained by solving the optimization problem. In many cases, to a first approximation, if the model is correct, the statistic that is the optimal solution has an approximately multivariate normal distribution whose mean equals the true mean and whose variances and covariances involve the second partial derivatives of the criterion function. Elliptical confidence regions for the parameter vector can be based on this approximation. This method of approximating the distribution of statistics will be referred to as asymptotic normality of parameter estimates. More accurate confidence regions can be obtained by a technique called inverting the criterion function, but they are computationally much more difficult. With either approach, simulations are useful to calibrate the approach (i.e., to improve the accuracy of coverage probabilities).

Methods of estimation are discussed further in Appendix B, which illustrates the calculation of criterion functions using the senior (age 65 or older) citizen drinking water percentile dataset.

2.2.8 Recommended Methods of Estimation and Discussion

MLE is the single most credible and most widely applied method and, therefore, is the method chosen for estimating exposure factor distributions. Caution is needed in the use of the MLE because many of its touted virtues depend strongly on the assumption that the model is true. For instance, if the model is correct, then the MLE converges to the correct value of the parameter as the sample size grows larger. On the other hand, if the true model is gamma or Weibull, but is assumed to be lognormal, then the MLE of the assumed lognormal mean converges to something other than the true mean. In addition, the common assumption that the variance of the MLE is given by the expected negative second partial derivatives of the log-likelihood function evaluated at the MLE will often lead to underestimation of the variance. Generally speaking, even if the MLE is used as the parameter estimate, consideration should be given to using other (regression or chi-square) methods to obtain variance estimates that are robust to model violations and give approximately unbiased variance estimates, even if the model is wrong.

2.3 Methods of Assessing Statistical Goodness-of-Fit (GOF)

# P-P plots, Q-Q plots, and percent error plots
# Likelihood ratio tests (LRTs) of fit versus a more general model
# F tests of fit versus a more general model
# Pearson chi-square tests of absolute fit
# Tests of absolute fit based on distances between distribution functions

GOF tests are tests of the null hypothesis that the assumed family of models is correct. As is evident from the discussion below, there is a natural correspondence between methods of estimation and methods of testing GOF. This stems from the fact that most of the criteria functions that drive the estimation process actually represent a type of fit to the data.

P-P plots and Q-Q plots, as well as GOF tests based on Pearson's chi-square and the empirical distribution function (EDF), are discussed and applied in Law and Kelton (1991).

2.3.1 P-P Plots, Q-Q Plots, and Percent Error Plots

P-P (probability-probability) plots, Q-Q (quantile-quantile) plots, and percent error plots are commonly used graphical displays that are applicable to models fit to raw data or to percentile data. These provide informal graphical aids for evaluating fit. P-P plots are made by plotting model-based estimates of probability on the vertical axis versus nominal probability on the horizontal axis. Both axes therefore go from 0 to 1. Q-Q plots show the model-based quantile estimates on the vertical axis versus empirical quantiles (Xp values) on the horizontal axis. Although P-P plots and Q-Q plots are informative, their regions of interest are near the main diagonal, and most of the plot field is blank. Percent error plots convey the same information but magnify the regions of interest by referring to a horizontal line instead of a diagonal line. Percent error probability plots are simply plots of ($\hat{P}$-P)/P versus P, where $\hat{P}$ denotes a model-based probability and P is an empirical or nominal probability. Percent error plots are defined analogously for quantiles.

P-P plots, Q-Q plots, and percent error plots do not take into consideration the number of estimated model parameters. Accordingly, they can be misleading if used to compare models with different numbers of parameters. A valid comparison of models requires the use of GOF statistics or *p*-values that take into account the number of estimated parameters.

2.3.2 Relative and Absolute Tests of Model Fit

Of the four test-based methods of assessing fit, two (LRTs and F tests) are tests of relative fit, and two (Pearson chi-square tests and tests based on EDF statistics) are tests of absolute fit. Relative tests of GOF are conducted by comparing one model (model 1) against another more general model (model 2) that contains the first model as a special case. Model 2 has more parameters than model 1, and model 1 is obtained by setting certain parameters of model 2 to fixed values. If model 2 is itself inadequate, little is gained by establishing the adequacy of model 1 relative to model 2. However, if model 1 is rejected relative to model 2, then model 2 improves the fit relative to model 1.

Tests of absolute fit of the model to the data are done without reference to any particular alternative model. Hence, they are more general than relative tests, because they do not require specification of a more general model.

2.3.3 Likelihood Ratio Test of Fit Versus a More General Model

LRTs are a natural companion to MLE because the two models are evaluated at their respective MLEs. The log-likelihood ratio is calculated by LR = -2*log(maxlik1/maxlik2), where maxlik1 (maxlik2) is the maximized log likelihood for model 1 (model 2). The GOF *p*-value usually is calculated by assuming the likelihood ratio has a chi-square distribution with degrees of freedom (df) given by the difference in dimensionality of the two parameter spaces. For example, the generalized gamma model contains the gamma, lognormal, and Weibull models as special cases, and allows for LRTs of the relative adequacy of these two-parameter models. In this case, df=3-2=1, since the generalized gamma has three parameters and the other models have two parameters.

One virtue of LRTs is the accuracy or reliability of their *p*-values, even for relatively small sample sizes. Generally, the performance of LRTs is much better than that of tests based on asymptotic normality assumptions (Barndorff-Nielsen and Cox, 1994).

2.3.4 F Test of Fit Versus a More General Model

F tests in nonlinear regression or WLS contexts provide another method of judging the adequacy of one model relative to a more general model. The F statistic is calculated as F=[(WSSE1-WSSE2)/(np2-np1)]/[WSSE2/(n-np2)], where WSSE1 and WSSE2 are the weighted sums of squares of errors for models 1 and 2, np1 and np2 are the number of parameters for models 1 and 2, and n is the number of data points. GOF *p*-values are calculated by assuming that F has an F distribution with df_1=np1-np2 and df_2=n-np2 degrees of freedom. This test can be used for linear or nonlinear models (Jennrich and Ralston, 1979).

2.3.5 Pearson Chi-Square GOF Test

The Pearson chi-square and EDF-based GOF tests are tests of absolute fit of the model to the data, without reference to any particular alternative model. Hence, they are more general than LRTs, because they do not require specification of a more general model but only compare a fitted model against the data.

The chi-square test is the simplest and most widely used of the absolute fit methods. The calculation of the summary chi-square value is described in Section 2.2. This chi-square calculation can be done using either the MLE or the MCS estimator to obtain the expected counts. Usually, the MLE is used, even though the MCS estimate minimizes the chi-square statistic. GOF *p*-values are calculated by assuming a chi-square distribution with df=c-np degrees of freedom, where c is the number of categories, and np is the number of model parameters. (Actually, the question of how many degrees of freedom to attribute to the chi-square does not have a firm answer [Law and Kelton, 1991, pages 384-385].)

2.3.6 GOF Tests Based on the EDF

Among absolute GOF tests, the chi-square test suffers from rather low power. Generally, tests based on the EDF are more powerful (Stephens, 1974). EDF tests involve generalized distances between the EDF and a theoretical CDF whose parameters have been estimated, usually by maximum likelihood. EDF tests based on Anderson-Darling (AD), Cramer-von Mises (CvM) and Kolmogorov-Smirnov (KS) distances are available. Although these EDF tests are more powerful than the chi-square test, their associated distribution theory is much more complex than that of the chi-square test. Tabulated approximations for AD and KS tests based on simulation studies for gamma, lognormal, and Weibull distributions are contained in D'Agostino and Stephens (1986). However, these do not easily adapt to the generalized F model, to censored data, or to models with point masses at zero. (Bootstrapping the test statistic is an option.) Despite its low power, the chi-square test is the most broadly applicable GOF test across distributional types.

2.3.7 Recommended Methods for Assessing Statistical GOF and Discussion

If raw data or several percentiles are available, P-P, Q-Q, and percent error plots are recommended as graphical aids in evaluation of GOF. A situation may arise where GOF tests indicate that one of the more general models fits considerably better than any of the two-parameter models. P-P, Q-Q, and percent error plots may be adequate for deciding which two-parameter model to use for a given exposure factor, but they may not lead to the right decision if the question is whether the best fitting two-parameter model is an adequate summary relative to a more general model with more than two parameters. Unlike GOF tests, these plots do not account for the number of estimated model parameters.

Another way to address this question is at the level of the overall risk assessment (RA), by sensitivity analysis. If two RAs are done, one with the best fitting two-parameter models, another with the absolute best fitting models, and negligible differences between bottom line measures of risk are obtained, then use of the simpler models is justified.

The chi-square and LRT GOF tests are recommended here because of their broad applicability and ease of use. If raw data are available, the AD GOF test for gamma, lognormal, and Weibull distributions may be used with tables in D'Agostino and Stephens (1986).

2.4 Methods of Obtaining Distributions for Parameter Uncertainty

# Method 1: Asymptotic normality of parameter estimates
# Method 2: Bootstrapping
# Method 3: Simulation from the normalized likelihood
# Method 4: Meta-analysis to combine multiple sources or studies

The first three methods for obtaining distributions of parameter uncertainty pertain to analyses of individual studies or datasets. Meta-analysis is used to combine results from two or more studies. Section 5 contains an example of meta-analysis for inhalation rates.

If we select and recommend a specific distribution that fits best to a given set of data, we neglect two kinds of uncertainty: uncertainty as to the type of model, and uncertainty in the numeric values of the model's parameters. A simple parametric model is not expected to capture a complex real-world situation exactly. Issues related to these two types of uncertainty are discussed in *Guiding Principles for Monte Carlo Analysis* (U.S. EPA, 1997b) and in Section 6.

2.4.1 Model Uncertainty

Regarding model uncertainty, three cases may be distinguished. In each case, it is assumed that several models have been fit to the available data, for example, the generalized gamma, gamma, lognormal, and Weibull.

# Case 1. One model fits adequately, and the other models are rejected. In this case, model uncertainty seems negligible, and the uniquely qualified model can be used for risk assessment.

# Case 2. All of the models are rejected, but one fits better than the others. If a model that fits cannot be found, then obviously model uncertainty is present. Nonetheless, one might work with the best fitting of the models tried, if the approximation is good enough. To give some indication of the effect of model uncertainty in risk assessment, the empirical distribution also might be included, in addition to the best fitting parametric model. Alternatively, some risk assessors might prefer to use the empirical distribution as the best guess distribution.

# Case 3. There is a virtual tie among two or more models. In this case, all of the viable models could be used for risk assessment.

In a sense, the distinction between the three cases is illusory, because the textbook distributions are conceded to be approximations in every case.

2.4.2 Parameter Uncertainty

Regarding parameter uncertainty, the fifth principle in *Guiding Principles for Monte Carlo Analysis* (U.S. EPA, 1997b) specifies that "for both the input and output distributions, variability and uncertainty are to be differentiated." The structurally sound approach of Rai et al. (1996) is followed: "Each variable is assumed to follow a distribution with one or more parameters reflecting population variability; uncertainty in the value of the variable is characterized by an appropriate distribution for the parameter values."

Four methods to obtain probability distributions for model parameters are discussed below.

2.4.2.1 Uncertainty Analysis Based on Asymptotic Normality of Parameter Estimates

Most parametric methods of statistical analysis can provide estimates of parameters as well as estimates of their variances and covariances. In the case of two-parameter models, this suggests that a certain bivariate normal distribution can be used for simulating the parameters, namely, the one with the estimated means and covariance structure. More generally, a multivariate normal distribution can be used. Caution must be exercised in the use of this approximate method that requires a large sample. It is difficult to provide simple guidance on how large a sample is required. The answer depends on specifics of the population distribution.

2.4.2.2 Uncertainty Analysis Based on Bootstrapping

The bootstrap method would generate many (e.g., 1,000) random samples of the same size as the original sample, drawn with replacement from the estimated ("best guess") distribution. Then, the modeling process would be applied to each such sample, resulting in an empirical distribution of estimated parameter values. This could be summarized as a data file with 1,000 records, each containing one set of parameter values. The risk assessor could sample at random from this list to obtain parameter values.

2.4.2.3 Uncertainty Analysis Based on the Normalized Likelihood

This method would normalize the likelihood function so it integrates to one over the parameter space. This normalized likelihood can be used as a probability distribution for the parameters. This method can be approximated by using a fine grid in the parameter space. The likelihood is evaluated at each grid point and divided by the sum of the likelihoods at all the grid points to obtain discrete probabilities. This discrete distribution can be sampled in proportion to these probabilities to obtain parameter vectors.

Methods 2 and 3 are much more computationally intensive than method 1. The risk assessor would not be expected to conduct the bootstrapping or likelihood normalization. Rather, the risk assessor could be provided with the appropriate data files for sampling.

It should be recognized that if the uncertainty distribution is inferred from a single study, then the treatment of uncertainty may be superficial and tend to neglect major portions of parameter uncertainty (Hattis and Burmaster, 1994, discussed further below). This is the rationale for the fourth method, based on meta-analysis.

2.4.2.4 Uncertainty Based on Meta-Analysis

Meta-analysis (discussed in Section 2.2.5) is a technique for synthesizing results from multiple studies. As part of a meta-analysis, it may be possible to obtain estimates of precision of the meta-estimates. These may be highly dependent on model assumptions. Meta-analysis is applied to estimate distributions of daily inhalation rates in Section 5.

Meta-analysis could be complicated by the fact that different types of probability models seem to be required for different studies. However, in many cases, it may be possible to proceed on the basis of the first two moments (mean and standard deviation), as in Section 5 on inhalation rates.

2.4.3 Recommended Method for Uncertainty and Discussion

The first of the four methods, based on asymptotic normality, is recommended for individual studies. The first method is simplest to apply because the required statistics are provided routinely by most methods of statistical analysis.

As described below, it would be possible to summarize each risk factor by providing two distributions, one that neglects uncertainty and one that incorporates uncertainty. An uncertainty distribution could then be obtained as the deconvolution of these two distributions. By conducting two risk assessments—using distributions neglecting uncertainty and using distributions incorporating uncertainty—variability and uncertainty could be differentiated.

The first distribution would be the one selected as providing the best fit to the available data. It is specified by identifying the appropriate type of distribution (e.g., gamma) and assigning the values for its parameters (e.g., the MLEs).

The second distribution would embody uncertainty in the model parameters, as well as population variability. It would be obtained by repeating a two-step simulation process many times and then summarizing, perhaps via additional modeling, the simulated data resulting from the two-step process. The two-step process involves first generating parameter values by sampling from the distribution representing parameter uncertainty, then generating a value for the variable of interest from the specified population distribution. This two-step process would be repeated many times (e.g., 10,000). Finally, the models can be fit to this simulated data to arrive at a best fitting distribution reflecting uncertainty.

Unfortunately, this approach is not adequate for all purposes (Paul White, statistician, Office of Research and Development, U.S. EPA, personal communication, Sept. 12, 1997). Interest in risk assessment typically centers on certain key parameters of the risk distribution, such as the mean and 95th percentile of the overall distribution of risk. Hence, to address uncertainty in a meaningful way in the context of the overall risk assessment requires that a distribution for such parameters be available. This implies that information on the distribution of the model parameters for each risk factor be provided. The risk assessor then can use these uncertainty parameter distributions to empirically generate distributions for the risk distribution parameter.

For example, a total of 10,000 simulations can employ an outer loop to generate 100 random sets of parameter values. For each set of parameter values, 100 population values are generated. For each step in the outer loop, the distribution of aggregate risk is calculated. This results in an empirical distribution for any risk parameter of interest, such as 95th percentile of risk or mean risk.

It is important to realize that there may be major neglected uncertainties beyond those that can be estimated from a single study. "The application of standard statistical techniques to a single dataset will nearly always reveal only a trivial proportion of the overall uncertainty" (Hattis and Burmaster, 1994). Each study reported in the scientific literature contains its own unique types of bias. These biases may be impossible to ascertain or estimate. In this case, the biases may be ascribed to randomness whose variance is estimated by meta-analyses that pool results across multiple studies.

2.5 System Output (Summary of Reported Statistics)

The most important summaries will be:

# Recommended type of model
# Estimated distribution for model parameters

Also reported would be variables identifying the data used for analysis, such as EFH table numbers, and the following statistics for each of the fitted models:

# Parameter estimates
# Parameter standard errors
# Asymptotic correlations between parameters
# Values of GOF statistics and associated *p*-values