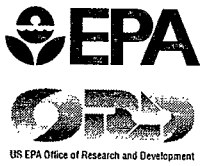
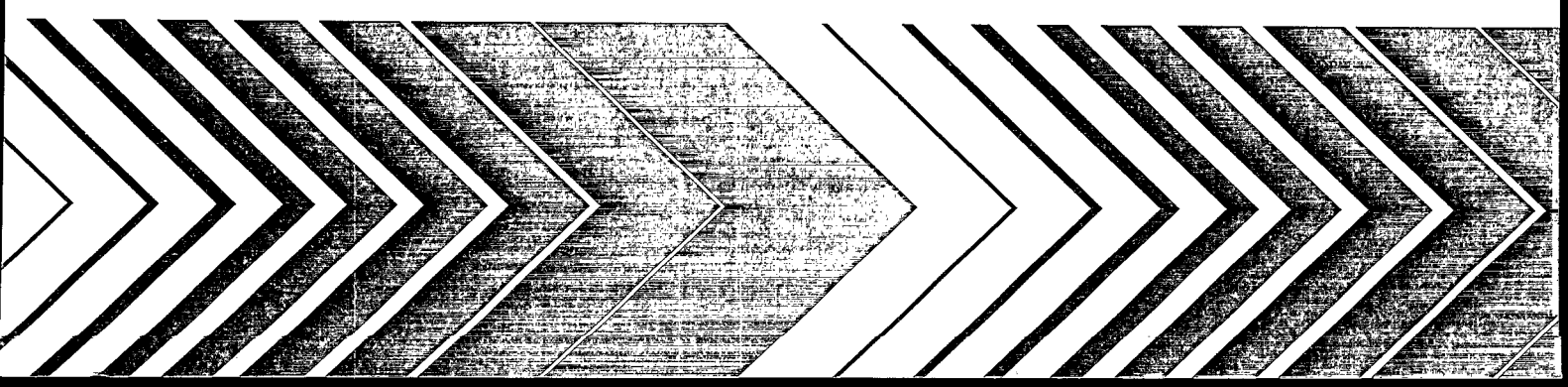
United States
Environmental Protection
Agency

Office of Research and
Development
Washington DC 20460

EPA/600/R-00/058
July 2000
www.epa.gov/ncea



Options for Development of Parametric Probability Distributions for Exposure Factors



EPA/600/R-00/058
July 2000
www.epa.gov/ncea

Options for Development of Parametric Probability Distributions for Exposure Factors

National Center for Environmental Assessment-Washington Office
Office of Research and Development
U.S. Environmental Protection Agency
Washington, DC

DISCLAIMER

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

CONTENTS

CONTENTS (continued)

CONTENTS (continued)

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AD	Anderson Darling
BMR	basal metabolic rate (MJ/day)
CDF	cumulative distribution function
CV	coefficient of variance
CvM	Cramer-von Mises
df	degrees of freedom
EDF	empirical distribution function
EFH	Exposure Factors Handbook
GOF	goodness-of-fit
H	oxygen uptake factor, the volume of oxygen (at standard temperature and pressure, dry air) consumed in the production of 1 MJ energy expended (m3/MJ)
iid	identically and independently distributed
KS	Kolmogorov-Smirnov
LRT	likelihood ratio test
MAAPE	minimized average percent error
MCS	minimum chi-square
MDE	minimum distance estimation
ML	maximum likelihood
MLE	maximum likelihood estimator
MOM	method of moments
NFCS	National Food Consumption Survey
NHANES	National Health and Nutrition Examination Survey

LIST OF ABBREVIATIONS (continued)

OLS	ordinary least squares
PAR	population at risk
PDF	probability density function
RA	risk assessment
WLS	weighted least squares
WSE	weighted sum of squares of errors

PREFACE

The National Center for Environmental Assessment (NCEA)–Washington Office within EPA's Office of Research and Development (ORD) has prepared this document in response to requests from users of the Exposure Factors Handbook (EPA/600/P-95/002Fa-Fc, August 1977) who expressed the need for assistance in using probabilistic methods in exposure assessments. This document summarizes procedures to fit distributions to selected data from the Exposure Factors Handbook.

AUTHORS AND REVIEWERS

The National Center for Environmental Assessment (NCEA)–Washington Office within EPA's Office of Research and Development was responsible for the preparation of this document. The original document was prepared by the Research Triangle Institute under EPA Contract No. 68D40091, Work Assignment No. 97-12. Jacqueline Moya of NCEA-Washington Office served as the EPA Work Assignment Manager, providing overall direction and coordination of the production effort.

AUTHORS

Lawrence Myers
Michael Riggs
Justin Lashley
Roy Whitmore

Research Triangle Institute
Research Triangle Park, NC

Jacqueline Moya
NCEA–Washington Office
Washington, DC

REVIEWERS

Preliminary drafts of this document received internal and external peer review. Reviewers included:

U.S. Environmental Protection Agency Reviewers:

Hans Allender	Office of Prevention, Pesticides, and Toxic Substances
Tim Barry	Office of Policy
Steven Chang	Office of Emergency and Remedial Response
Henry Kahn	Office of Water
Patrick Kennedy	Office of Prevention, Pesticides, and Toxic Substances

AUTHORS AND REVIEWERS (continued)

Steven Knott	Risk Assessment Forum, National Center for Environmental Assessment
Elizabeth H. Margosches	Office of Prevention, Pesticides, and Toxic Substances
Tom McCurdy	National Exposure Research Laboratory

External Reviewers:

Robert Fares	Environmental Standards Inc.
Paul Price	Ogden Environmental and Energy Services
Curtis Travis	Project Performance Corporation
Alan Stern	New Jersey Department of Environmental Protection
Shree Y. Whitaker	National Institute of Environmental Health Sciences

1

Introduction

The EPA Exposure Factors Handbook (EFH) was published in August 1997 by the National Center for Environmental Assessment of the Office of Research and Development (EPA/600/P-95/Fa, Fb, and Fc) (U.S. EPA, 1997a). Users of the Handbook have commented on the need to fit distributions to the data in the Handbook to assist them when applying probabilistic methods to exposure assessments.

This document summarizes a system of procedures to fit distributions to selected data from the EFH. It is nearly impossible to provide a single distribution that would serve all purposes. It is the responsibility of the assessor to determine if the data used to derive the distributions presented in this report are representative of the population to be assessed.

The system is based on EPA's *Guiding Principles for Monte Carlo Analysis* (U.S. EPA, 1997b). Three factors—drinking water, population mobility, and inhalation rates—are used as test cases. A plan for fitting distributions to other factors is currently under development.

EFH data summaries are taken from many different journal publications, technical reports, and databases. Only EFH tabulated data summaries were analyzed, and no attempt was made to obtain raw data from investigators. Since a variety of summaries are found in the EFH, it is somewhat of a challenge to define a comprehensive data analysis strategy that will cover all cases. Nonetheless, an attempt was made to ensure that the procedures used in the three test cases are fairly general and broadly applicable.

A statistical methodology was defined as a combination of (1) a dataset and its underlying experimental design, (2) a family of models, and (3) an approach to inference. The approach to inference itself may encompass a variety of activities (e.g., estimation, testing goodness-of-fit, testing other hypotheses, and construction of confidence regions). For present purposes, the approach to inference was limited to estimation, assessment of fit, and uncertainty analysis.

This section presents a review of important statistical concepts (Sections 1.1-1.5) and a skeletal summary of the recommended system (Section 1.6). A more detailed explanation of the system is provided in Section 2. Technical, mathematical, and statistical details were kept to a minimum. For instance, formulae for probability density functions, cumulative distribution functions, or means and variances of the different types of distribution are not presented. In addition the systems of equations that must be solved to obtain maximum likelihood and other types of estimates are not presented. Instead, references are given, and ideas are communicated intuitively. Appendices to this document contain some of the details. Appendix A contains a glossary and a list of abbreviations.

1.1 Review of Pertinent Statistical Theory and Concepts

A numeric event whose values change from one population member to the next is called a *random variable*. A random variable that takes only a finite number of values is called a *discrete random variable.* The number of carrots consumed in a day is a discrete random variable. By contrast, a *continuous random variable* can take on an infinite number of values over its range, that is, the total dry-weight of the carrots consumed in a day. However, in practice, the number of possible values for a continuous random variable will be limited by the precision of the instrument used to measure it. Because this report describes procedures for fitting theoretical distributions to continuous data, this review will be confined to the statistical properties of distributions of continuous random variables.

Samples of random variables often are summarized by their frequency distributions. A frequency distribution is a table or a graph (Figure 1-1a) that displays the way in which the frequencies (i.e., counts) of members of the sample are distributed among the values that they take on. The relative frequency distribution (Figure 1-1b) can be calculated by dividing each count by the total sample size. If the counts are large enough, it is often possible to summarize the relative distribution with a mathematical expression called the *probability density function* (PDF). The PDF predicts the relative frequency as a function of the values of the random variable and one or more constraining variables, called model parameters, that can be estimated from the sample data. Continuous distributions whose PDFs can be so defined are called *parametric continuous distributions*. In Figure 1-1b, the plot of a PDF for a normal distribution is superimposed on the relative frequency distribution of the continuous random variable, X, from which it was computed. The mathematical expression for the normal PDF is

$$\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\left(\frac{1}{2\sigma^2}\right)(x-\mu)^2\right].$$

In this example, the two parameters of the PDF are the population mean $\mu = 5.0$ and the population standard deviation $\sigma=1.58$. The area under any PDF curve is 1.0 and represents the probability of observing a value of x between the population minimum and maximum. The probability that X will be contained in some interval [X=a, X=b] can be calculated simply by integrating the PDF from a to b:

$$\Pr[a<X<b] = \int_{x=a}^{b} \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\left(\frac{1}{2\sigma^2}\right)(x-\mu)^2\right]\, dx.$$

It follows that the probability that X equals any particular value x is zero.

In epidemiology, many situations arise in which a measurable fraction of the study population has not been exposed to the risk factor of interest. For example, the distribution of tap water consumption by infants on any given day would be expected to have a relatively large number of zero values. This poses a problem to the risk modeler who attempts to fit a parametric PDF because the associated models all predict an infinitesimal probability for any point value of X, including zero. One compromise is to ignore the zeros and fit the model to the infant subpopulation that actually consumes tap water. Obviously, this will not be helpful to the modeler who needs to model the entire population. The solution is to fit a composite PDF model to the data such that the unexposed subpopulation is assigned a fixed point-probability of being unexposed while the exposed population is modeled with one of the usual PDF families. Because such models allow a positive probability density at X=0, they are referred to as PDFs with a *point mass at zero*. An example of the plot of a lognormal exposure distribution with a 0.06 point mass at zero (i.e., 6% unexposed) is illustrated in Figure 1-3f. The mathematical expression for its composite PDF is

$$f(x) = \begin{cases} 0.06 & \text{if } x=0 \\ \dfrac{1}{x\sigma\sqrt{2\pi}}\exp-\left(\dfrac{(\log(x)-\mu)^2}{2\sigma^2}\right) & \text{if } x > 0 \end{cases}$$

Another function often used to describe parametric distributions is the *cumulative distribution function* (CDF). The CDF is the probability that $X \leq x_i$ for all x_i in the population. Many of the more commonly used nonparametric tests of differences in the distributions of continuous random variables evaluate hypotheses about the CDFs. Plots of the PDF and CDF of a lognormal distribution are illustrated in Figures 1-2a and 1-2b. PDFs from five additional families of continuous parametric distributions are illustrated in Figsures 1-3a–1-3f. These and other families considered in this report differ from the normal distribution in that they are defined only for positive values. Because its domain includes negative values, the normal distribution is not useful for modeling environmental exposure factors. However, the log-transformations of many exposure factors and spatially aggregated environmental variables are approximately normally distributed. For this reason, the lognormal is frequently employed to model environmental and epidemiologic data.

A thorough treatment of the various families of parametric continuous random distributions can be found in Johnson and Kotz (1970) or, in more concise form, in Evans et al. (1993). For any of these general families, an infinite number of distributions can be generated by varying the values of the parameters of the PDF (e.g., Figure 1-3a). However, regardless of the model, several methods are available for fitting a parametric PDF to a sample of continuous data values. The method employed throughout this report involves the fitting of PDFs by maximum likelihood estimation of the parameters of the PDF model. Maximum likelihood estimation is reviewed in Section 1.2. Brief discussions of some alternative parametric model fitting and estimation procedures are presented in Section 2. With modifications and some penalties, these same methods also can be used to fit PDFs to quantiles and/or other sample statistics (e.g., the mean and standard deviation). *Quantiles* are descriptive statistics whose q-1 values divide a sample or population of a random variable into q portions, such that each portion contains an equal proportion of the sample or population. For example, percentiles are obtained when q=100, deciles when q=10, and quartiles when q=4. The distributions reported in the EFH are summarized by the minimum, maximum, and 5^{th}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, and 95^{th} percentiles. Thus, all the examples presented in this report describe procedures for fitting parametric PDFs to these nine quantiles.

Although this report primarily concerns the fitting of parametric distributions to empirical data, it is important to note that alternative approaches can be used. Of particular importance are two methods of probability density estimation that do not require the a priori specification of an underlying parametric model. Both are based on the attributes of the observed sample *empirical distribution function* (EDF).

As its name implies, the EDF is the empirical counterpart to the theoretical parametric CDF; that is, the EDF is the probability that $X \leq x_i$ for all values of x in the *sample*. The EDF is the sum of the relative frequencies of all sample values of $X \leq x_i$. Its plot is a monotonically increasing step-function from zero to one (Figure 1-6a).

The first nonparametric method, kernel density estimation (Bowman and Azzalini, 1997), is an extremely flexible method for estimating a smoothed probability distribution from an EDF. Like the parametric approach, this method involves the fitting of a smooth curve to the relative frequency distribution. This is done by selecting an "optimal" nonparametric smoothing function. Several selection techniques are available, but most employ criteria that minimize the mean square error (e.g., ordinary least squares cross-validation). The second nonparametric method uses the EDF itself as the source of all probability density information. This approach is especially appropriate for large samples (e.g., $n \geq 1{,}000$) wherein it can be argued that there is sufficient information to make broad inferences regarding the population distribution. Both nonparametric methods have the advantage of providing density estimates without resorting to restrictive (and perhaps unrealistic) parametric assumptions. However, they are less portable than their parametric counterparts, that is, there is no well-studied reference distribution with known properties on which to rely. Also, their specification in risk assessment simulations is more difficult than parametric model specification. The specific EDF must be available to each investigator who wishes to apply it, and its properties must be independently investigated and verified.

A critical assumption for all the estimation methods so far discussed is that the sample observations are identically and independently distributed (the *"iid" assumption*). By "identically," we mean that all the sample members come from a population with a single PDF. "Independently" means that the random variable values are not correlated among members of the population. In multivariable risk assessment models there is an additional independence assumption, namely, that the values of the covariates are not correlated with one another. In fact, this often is not the case. For example, the distribution of dietary intakes for 8-year-old children may be composed of six components—water, vegetables, fruits, dairy products, meat, and fish—the relative amounts of which are correlated. Thus, children who eat large quantities of fruit and dairy products may eat relatively little meat compared with children who consume small amounts of dairy and fruit but large quantities of water. Depending on the nature of these correlations, the joint distributions for the six intake categories will differ among children. Multivariable mixtures of this kind are called *multivariate distributions*. Parametric

multivariate distribution models include correlations among their parameters and thus do not require independence. In contrast, the univariate models assume that the six intake PDFs are the same for all 8-year-olds, vegetarians and nonvegetarians alike. Although this may be unrealistic, in many cases (perhaps most), information on the multivariate correlation structure will not be available. Thus, the univariate approach may be the best option for the risk modeler. In such less than ideal situations, the univariate methods presented in this report may be quite useful. However, it should be understood that results based on such data must be interpreted with caution.

Statistical analyses should be consistent with the underlying study design. Many exposure factor data sets are from complex sample surveys. Proper analysis of survey data requires that weights and other design features such as clustering should be taken into account. The methods of inference that are used in this document can be easily adapted to complex survey data (Krieger and Pfeffermann, 1997). Survey data analysis software such as Research Triangle Institute's SUDAAN (Shah et al., 1997) can be used to obtain weighted percentiles and appropriate associated standard errors. This can be done for selected percentiles such as the nine deciles, or the entire weighted EDF can be estimated (along with standard errors appropriate to the EDF at each sample point). Finally, likelihood or minimum distance methods analogous to those applied in the elementary simple random sampling context can be used to estimate parametric distributions conforming as closely as possible to the survey-weighted percentiles or EDF, in a way that takes account of both the survey weights and the clustering.

So far, we have discussed methods for fitting PDF models to empirical data by estimating the appropriate parameters from the sample data. Having completed this step, the modeler is left with the question of whether the estimated model(s) actually fits the sample data. Figures 1-3a, 1-3e, and 1-3f illustrate a situation where this is not straightforward. Although the three PDFs are different, they have the same mean (20) and standard deviation (16) and very similar shapes. In fact, there are many models with mean=20 and standard deviation=16 that could be considered for a given set of data. Clearly some method of assessing the goodness-of-fit of each PDF model is required. Section 2.3 of this report summarizes several goodness-of-fit tests that evaluate the null hypothesis that the EDF and model CDF are equal. In addition, three graphic procedures for visually assessing the CDF to EDF fit are introduced (Section 1.4). Criteria based on joint consideration of the goodness-of-fit tests and EDF graphics can be used to resolve the problem of model selection that is exemplified by the similarities of Figures 1-3a, 1-3e, and 1-3f. These criteria are discussed in Section 2.3.

1.2 Maximum Likelihood Estimation

Given a set of observed data, it is often of interest to develop statistical models to investigate the underlying mechanisms that generated the data (causal models) and/or to predict future distributions of the same variable (prognostic or predictive models). Usually there will be more than one model that reasonably can be considered for the process or system under investigation. As a first step in determining which of the models best fits the data, it is necessary to estimate the values of the parameters of each hypothesized model. Several methods are available; among the most commonly used are the method of moments (MOM), ordinary least squares (OLS), weighted least squares (WLS), and maximum likelihood (ML). These methods and others have specific advantages and disadvantages. However, a preponderance of statistical theory, research, and experience indicate that estimates obtained by ML have minimum bias and variability relative to competing estimators in a very broad range of contexts (Efron, 1982). For this reason and others that are explained later, we have chosen to rely primarily on maximum likelihood estimators (MLEs) in developing the methodology of Sections 2 and 3. Herein, we present a brief introduction to MLE.

Suppose we obtain a sample of nine fish from a pond and we want to estimate the prevalence of *aeromonas* infection (red sore disease) among fish in the pond. Because each fish must be counted as either infected or uninfected, the binomial probability model is an immediate candidate for modeling the prevalence. The binomial probability function is

$$\Pr(y = Y \mid p) = \binom{n}{y} p^{y} (1-p)^{n-y}$$

where y=number of fish in the sample with red sore lesions

n=the sample size (nine fish)

p=the probability of infection ($0 \le p \le 1$)

Clearly y and n are obtained directly from the data, leaving only p to be estimated. Suppose further that we hypothesize three possible prevalence rates 0.20, 0.50, or 0.80; we can now construct a table of the predicted probabilities of observing y=0,1,2...9 infected fish in a sample, given the binomial model and each of the three hypothesized values of p. The predicted probabilities in Table 1-1 are obtained by substituting these values into the binomial PDF. The results indicate that P=0.20 yields the highest likelihoods for samples with three or fewer infected fish, P=0.50 for samples with four or five

infected fish, and P=0.80 for samples with more than five infected fish. This example demonstrates that the value of the MLE depends on the observed data. Accordingly, we define an MLE as that parameter estimate that yields the largest likelihood for a given set of data.

For illustrative purposes, we specified three candidate parameter values a priori. In practice, one usually specifies only the model and then estimates the parameters from the observed data. For example, suppose we have four infected fish in a sample of nine. What is the MLE of P? We can obtain the MLE by trial and error simply by varying P from 0 to 1.0 in very small increments (e.g., 0.001) with y=4 and n=9, substituting them into the binomial PDF, and plotting the resulting likelihoods versus P. To further illustrate the data-dependent nature of the MLE, we will repeat this exercise for y=2 and y=8. The results are plotted in Figure 1-4. By inspection, we see that for samples with two, four, and eight infections, the corresponding MLs occur at P=0.22 (2/9), P=0.44 (4/9), and P=0.89 (8/9), which are of course the observed proportions of infection. In fact, for any sample, the MLE of the binomial parameter P will always be the sample proportion, y/n.

The preceding simple exercise illustrates the essential steps in computing an MLE:

- Obtain some data.
- Specify a model.
- Compute the likelihoods.
- Find the value of the parameter(s) that maximizes the likelihood.

In this example, we estimated a single parameter by eyeballing the maximum of the plot of the likelihood. However, most applied statistical problems require the simultaneous estimation of multiple model parameters. For such cases, the maximum of the likelihood curve for each parameter must be obtained by application of methods from differential calculus. Details of the mathematics are available in most introductory mathematical statistics texts (e.g., Mendenhall et al., 1990); however, risk assessors may find the more elementary (but complete) treatment of MLE by Kleinbaum et al. (1988) to be more understandable.

Because many multivariate likelihood functions are nonlinear, closed-form solutions to the differential equations often will not exist. Instead, computationally intensive iterative algorithms will have to be applied to get the desired parameter MLEs. These algorithms are widely available in statistical software packages (e.g., SAS and SPLUS) and execute quite rapidly on modern computers. The same algorithms can be used to obtain the MLE of the variance-covariance matrix of the estimated

model parameters. These estimates are crucial for statistical tests on the parameters and for estimates of parameter uncertainty. For a model with P parameters, the associated variance-covariance matrix will be PxP with the variance estimates of the parameters on the diagonal and the corresponding parameter covariance estimates in the off-diagonal positions. For example, the variance-covariance matrix of a model with three MLE parameters, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ is:

$$\begin{bmatrix} Var\,\hat{\beta}_0 & Cov(\hat{\beta}_0, \hat{\beta}_1) & Cov(\hat{\beta}_0, \hat{\beta}_2) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var\,\hat{\beta}_1 & Cov(\hat{\beta}_1, \hat{\beta}_2) \\ Cov(\hat{\beta}_0, \hat{\beta}_2) & Cov(\hat{\beta}_1, \hat{\beta}_2) & Var\,\hat{\beta}_2 \end{bmatrix}$$

For models that require independence among their parameters (e.g., normal theory, Analysis of Variance [ANOVA], and regression), the covariance terms are assumed to be zero; however, other models (e.g., mixed model repeated measures ANOVA) permit nonzero correlations. In the case of the independence models, parameter tests and estimates depend only on the diagonal elements of the variance-covariance matrix. For all other models, the covariance terms must be taken into account when constructing statistical tests or confidence intervals. The MLE variance-covariance matrix is routinely computed by most statistical software packages that employ MLE.

One of the most useful and important properties of MLEs is that the ratio of the MLE to its standard error has a normal distribution with mean zero and standard deviation of 1.0.; that is,

$$\hat{\beta}_i / \sqrt{Var\,\hat{\beta}_i} \approx N(0,1).$$

Another way of saying this is that the ratio is a standard normal variate (Z-score). Therefore, comparison of the ratio to the Z distribution provides a test of H_0:β_i=0. Alternatively, it can be shown that

$$\hat{\beta}_i^2 / Var\,\hat{\beta}_i \approx \chi^2$$

with n-(1+p) degrees of freedom—[where n=the size of the sample from which MLE(β_i) was computed and p=number of ML estimates computed from the sample]—permitting one to test the same hypothesis against the chi-square distribution. These relationships lead directly to the formation of 1-α% confidence intervals about $\hat{\beta}_i$:

$$\hat{\beta}_i \pm Z_{1-(\alpha/2)} \sqrt{Var\,\hat{\beta}_i}$$

where $Z_{1-(\alpha/2)}$ is the Z-score associated with a probability of 1-α; for a 95% confidence interval, α=0.05 and Z=1.96.

The width of the confidence interval is indicative of the degree of uncertainty associated with the MLE. The narrower the confidence interval, the more certain we are of the estimate.

The properties of minimal bias and variability, as well as that of normality, can be assured only when the MLE is based on "large samples" of data. Optimally "large" means $n \geq 30$. While $20 \leq n < 30$ will often provide reasonably good MLEs, MLEs computed from samples of $10 \leq n < 20$ should be viewed with caution and those based on $n < 10$ should be avoided altogether. This is because the sampling distribution of an MLE becomes less normal, biased, and more variable as n approaches zero. Conversely, the distribution tends to normality as n gets increasingly large. This tendency is called asymptotic normality.

The relationship among the MLEs, their standard errors, and the chi-square distribution is the basis for an extremely useful and versatile class of statistical tests called likelihood ratio tests (LRTs). An LRT statistic is formed from the ratio of the likelihoods associated with two MLEs. By definition, these are the maximum values of the likelihood of observing the given data points under the specified model. For example, the LRT formed between the binomial model MLEs associated with y=2 and y=4 in our fish sampling problem would be the ratio of the infection likelihoods 0.306 and 0.260 (Figure 1-4). It can be shown that -2 times the log of the ratio of two such maximum likelihoods will be distributed as a chi-square with degrees of freedom equal to the number of ML parameters of the denominator likelihood minus those of the numerator likelihood. In the example just described, the numerator and the denominator have the same number of parameters (1), so the chi-square test cannot be carried out.

LRTs are used primarily for choosing among hierarchical multivariate models. Consider a model for the random variable y for which three MLE parameters, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$, have been estimated. A fundamental tenet of mathematical modeling is that parsimonious models are the most efficient models. Thus, we would like to determine whether two- or single-parameter versions of our model would do as good a job of describing our data as the full three-parameter model. This is done by forming a series of LRTs, each with the likelihood of the model with the lesser number of parameters as its numerator. To test whether the full model performs better than a model containing only the first two parameters, we would form the following LRT statistic:

$$-2\ln\left[\frac{L\left(y|\hat{\beta}_0,\hat{\beta}_1\right)}{L\left(y|\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2\right)}\right]$$

where $L\left(y|\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2\right)$ = the likelihood associated with the full model

$L\left(y|\hat{\beta}_0,\hat{\beta}_1\right)$ = the likelihood associated with the two-parameter model.

The difference between the number of parameters in the denominator and numerator models is 3-2=1. Thus, the LRT can be compared to a chi-square with one degree of freedom. This LRT evaluates H_0: $\beta_2=0$; rejection at the specified α provides evidence that the three-parameter model is necessary. Acceptance of H_0 would provide evidence in favor of the alternative two-parameter model. Tests comparing three-parameter or two-parameter models with each other or with any of the three possible one-parameter models can be formed by substituting the appropriate likelihoods into the above expression and comparing them to the appropriate chi-square distribution.

The LRT depends on the fact that

$$L\left(y|\hat{\beta}_0\right) < L\left(y|\hat{\beta}_0,\hat{\beta}_1\right) < L\left(y|\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2\right).$$

This relationship will exist only when the various models are formed by the deletion of one or more parameters from the full model. Series of models that are so constructed are called hierarchical or nested models. LRTs formed between models that are not hierarchical will not necessarily follow a chi-square under H_0 and are therefore invalid. By definition, two models with the same number of parameters are not hierarchical; thus, the LRT that we attempted to form earlier from the two binomial models did not lead to a valid chi-square test.

In summary, MLEs:

- Provide estimates that are the most likely (consistent) given the observed data
- Have minimum bias and variance and are asymptotically normal for n≥30
- Allow easy estimation of parameter uncertainty
- Provide a flexible means of model fitting through LRTs

1.3 Probability Models

The parametric distributional models described in this report are mathematical functions of continuous random variables and one or more model parameters. The numbers and kinds of parameters together with their functional form may be used to generate different families of exposure distributions. Sixteen such families are listed in Section 1.6; mathematical details are provided in Appendix D. Each family may be defined in terms of one or more of the following types of parameters:

- *Location* parameters define a point either at the center or in the upper and/or lower tails relative to which all other points in the distribution are located. For example, the mean (μ) marks the center of the distribution of a normal random variable while the degrees of freedom (df_1, df_2) mark the tails of an F-distributed variable. Thus, normally distributed variables with different means are shifted with respect to one another, and two F-distributed variables with different degrees of freedom will have different densities in their respective tails.

- *Scale* parameters define the amount of spread or variability in the distributions of continuous random variables. For example, if we have two normal distributions with the same location ($\mu_1=\mu_1$) but with different sized variances ($\sigma_1<\sigma_2$), the one with the larger variance will tend to have more extreme values than the other, even though on average their values will not differ.

- *Shape* parameters are parameters that determine any remaining attributes of the shape of a probability and/or frequency distribution that are not determined by either location and/or scale parameters. These may include, but are not limited to, skewness and kurtosis.

Environmental distributions tend to concentrate on the nonnegative real numbers with a long right tail and are often approximated using lognormal, gamma, or Weibull two-parameter models. The one-parameter exponential model, a special case of the gamma and Weibull models, is occasionally useful. In the majority of cases, however, two or more parameters are required to achieve adequate fit. The generalized (power-transformed) gamma distribution is a three-parameter model that includes the gamma, lognormal and Weibull models as special cases (Kalbfleisch and Prentice, 1980). Because of the popularity of these two-parameter models, the generalized gamma distribution is a particularly important

generalization. The SAS Lifereg procedure will fit regression models based on the generalized gamma model. The generalized gamma is obtained by simply raising a two-parameter gamma random variable to a positive power.

An even more general model, which includes most of those encountered in practice as special cases, is the four-parameter generalized F distribution. An F random variable is the ratio of two independent gamma or chi-square random variables. The generalized F random variable is a power-transformed F, that is, it is obtained by raising an F variable to some power. The generalized F distribution is a four-parameter model that includes the generalized gamma model as a special case, as well as the two-parameter log-logistic model and the three-parameter Burr and generalized Gumbel distributions (Kalbfleisch and Prentice, 1980). Appendix D contains formulae for probability density functions, cumulative distribution functions, and moments for the generalized F distribution and many of its special cases.

Our treatment of the generalized F distribution is not intended to be exhaustive. Excellent sources of additional information are Chapter 2 and Section 3.9 of Kalbfleisch and Prentice (1980) and the classic books on distributions by Johnson and Kotz (1970).

Kalbfleisch and Prentice (1980) show graphically how various special cases of the generalized F can be envisioned in a two-dimensional graph, with the horizontal and vertical axes representing the numerator and denominator degrees of freedom (df_1 and df_2) for the F random variable. For instance, the log-logistic model has $df_1=df_2=2$, the generalized gamma distribution is obtained by letting df_2 approach infinity, and the lognormal model is obtained by letting both degrees of freedom tend to infinity.

We have found that the most useful cases of the generalized F are those listed below, with number of parameters in parentheses.

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

A further generalization that is sometimes useful is to adjoin a point mass at zero to account for the possibility that some population members are not exposed. This increases the number of parameters by one.

One question that is sometimes raised is whether the use of the generalized F distribution constitutes overfitting. According to Norman Lloyd Johnson, the world's foremost expert on parameter probability distributions, "fitting a maximum of four parameters gives a reasonably effective approximation" (Johnson, 1978). A more complete reply to the overfitting question is as follows. Suppose we are in the fortunate situation where we have a few hundred or a few thousand observations, and we want to fit a smooth curve model to the empirical distribution of the data. There is no reason why nature should have dictated that a mere two parameters would account for the behavior in both tails as well as the mid-range. In such a situation of extensive data, we find it perfectly reasonable to allocate two parameters to the lower tail and two other parameters to the upper tail. But this is precisely how the generalized F works: as the population variable x decreases to zero, the generalized F probability density function behaves like a power function $a_1 x^{b_1}$; as the population variable x increases to infinity, the generalized F probability density function behaves like a different power function $a_2 x^{b_2}$. The generalized F is as simple and natural as this: it allows the two tails to be modeled independently, allocating a power function with two parameters for each. In fact, a need for six-parameter models is clear enough, allocating two more parameters to the mid-range.

It is important to emphasize that all of the distributions described in this report are just special cases of the generalized F distribution, and they can be generated by setting one or more of the parameters of the generalized F to specific values such a 0, 1, or infinity. Thus, the sequence of 16 distributional families listed in Section 1.6 constitute a hierarchical set of models. This property allows us to apply the LRT methodology introduced in the previous section to select the "best" parametric model for a particular sample of data and motivated the development of most of the procedures described and implemented in this report.

1.4 Assessment of Goodness-of-Fit

The methods described in Section 1.2 allow optimal estimation of the parameters of any of the 16 candidate hierarchical models listed in Section 1.6. Once this is done, LRTs can be used to determine which of the models best fit the observed data. However, the LRT provides only a *relative* test of goodness-of-fit (GOF); it is entirely possible that the model with smallest log-likelihood p-value may be the best among a group of very poor competitors. Clearly, some method of assessing the *absolute* GOF is desirable.

The first task is to define a criterion for absolute GOF. Perhaps the simplest method is to subtract the observed data values from those predicted by the model with the fitted parameters. By definition, this difference will be near zero for models that closely fit the data. This approach is employed universally for evaluating the GOF of multiple linear regression and multiple logistic regression models. Residual (i.e., observed-predicted) plots are used to evaluate both fit and validity of model assumptions, while lack-of-fit and deviance tests are used to evaluate H_0: the regression model fits the data. In an analogous manner, both graphic and test-based methods can be applied to evaluate observed data values versus those predicted by a parametric probability model.

Unlike multiple regression models that predict a mean or a proportion, the probability models in Section 1.3 predict an entire population distribution. Thus, the GOF criteria must be applied to statistics that specify all the data points of interest. Accordingly, we employ methods that compare the EDF of the sample data to the fitted CDF of a specified parametric probability model. Because the EDF and CDF define explicitly the probabilities of every data point, they can be used to compare the observed sample with the type of sample that would be expected from the hypothesized distribution (Conover, 1980).

In this section, we introduce four graphical methods for comparing EDFs to CDFs and a GOF test of H_0: EDF=CDF, based on the chi-square distribution. Although several alternative GOF tests are described briefly in Section 2.7, we employ the chi-square GOF test and the four graphical methods almost exclusively for the evaluation of models described in this report.

We illustrate these techniques with the EFH data for tap water consumption by persons 65 years of age or older (n=2,541). The data were originally presented as percentile summaries (EFH Table 3-7) and are partially reproduced in Table 1-2 and in Table B-1 of Appendix B of this report. The first column of Table 1-2 lists the percentiles, and the second column lists the corresponding values of tap water consumption (mL/kg-day). Columns 3 and 4 are the actual and predicted proportions of the sample that are in the interval. For example, 4% of the sample consumed between 4.5 and 8.7 mL/kg of water per day versus the 4.777% predicted by the two-parameter gamma model. While the observed probabilities were computed from the EFH data, the predicted gamma probabilities were computed from the ML estimates of the gamma parameters (Table 1-2) using SAS software. The observed and expected numbers of people in each interval are, respectively, the product of the observed and predicted probabilities with 2,541 (n), the total sample size. Computation of the last column is explained later.

The observed probability distribution (EDF) and the predicted probability distribution (CDF) are computed by summing the observed and predicted probabilities. The simplest and most direct way to compare the two distribution functions is to overlay their plots (Figure 1-5a). The CDF is continuous for all possible data points, but the EDF is a step function with steps at each of the nine reported sample percentiles. These are the only points for which information is available; however, at these nine points, the CDF and EDF values agree very closely. The large sizes of the steps reflect the relative paucity of information carried by the nine sample percentiles. Had the raw data been available, the steps would have been more numerous, much smaller, and closer together.

An alternative, clearer way to compare the CDF with the EDF is illustrated in Figure 1-5b. This plot differs from the plots of the distribution functions in two respects. First, the observed values are replaced on the horizontal axis by CDF. Since both axes represent probability measures, this type of graph is called a probability-probability (P-P) plot. The diagonal line is the plot of the CDF against itself and corresponds to the line of equality with the CDF. The second difference is that only the left top corner of each step of the EDF is plotted (open circles). Because the EDF values are plotted against the CDF values, proximity of the circles to the diagonal is an indication of good fit to the model. Although this graph carries all the information of Figure 1-5a, it is much easier to interpret. Both figures provide evidence that the gamma model is a very good fit to the EFH sample data.

Figure 1-6a, a rescaled version of the probability (P) plot, is called a percent error probability plot. The vertical axis values are computed as:

$$\% \; Error \; = \; \frac{\hat{P}_i - P_i}{P_i}$$

where $\hat{P}_i$ = the CDF value in the i^{th} interval
P_i = EDF value in the i^{th} interval.

Plotting the proportionate deviation of the predicted from the observed versus the observed magnifies the deviations and permits comparison with the horizontal reference line corresponding to 0% difference. Based on this plot, it appears that lower values of tap water consumption deviate from the gamma model. However, the only really large deviation (-58%) is associated with the first percentile of tap water

consumption. This indicates that the model fails only in the lowest extreme of the consumption distribution; for all other data, the model appears to perform quite well.

The final graphical technique, called a quantile-quantile (Q-Q) plot, compares the observed quantiles of tap water consumption to those predicted by the gamma probability model. While the former were computed from the data (column 2, Table 1-2), the latter were obtained by programming the SAS RANGAM probability function. Obviously the scale of the two axes are different. However, this is simply a reflection of the units used to measure the observed data. In this case, the observed values are 100 times the predicted quantiles. Thus, the diagonal reference line marks the points where the observed values equal 100 times the predicted. The plotted points (open circles) mark the coordinates of the paired observed and predicted quantiles. Because the plotted points all lie very close to the diagonal, we may conclude that quantiles differ by not much more than the 100× scaling factor. This graph is further indication that the gamma model fits the data well.

Percent error P-plots, P-P plots, Q-Q plots, and percent error Q plots (the quantile equivalent of the percent error P-plots) are employed throughout this report to assess GOF. To improve readability, "Nominal P" and "Estimated P" are substituted, respectively, as axis labels for EDF and CDF.

The Pearson chi-square GOF statistic is computed as:

$$T = \sum_{i=1}^{c} \frac{(O_i - E_i)^2}{E_i}$$

where O_i = the observed frequency in the i^{th} interval
E_i = the expected frequency in the i^{th} interval
c = the number of intervals.

The c intervals are arbitrarily defined but should be constructed so that at least 80% of them have expected frequencies greater than five (Conover, 1980). For the data in Table 1-2, c=10 (the number of rows) and T is the sum of the values in the last column. The "Cell Chi-Sq." column contains the squared deviations of the observed from the model predictions. Small values of the cell chi-square indicate a good fit in the interval; large values indicate a lack of fit. For a model that provides a perfect fit to the data, the expected value of T is zero. Thus, small values of T indicate a good fit. Under the null hypothesis that the model fits the data, T will be distributed as χ^2 with c-(1+p) degrees of freedom, where

p is the number of estimated parameters in the fitted model. Since a two-parameter gamma model was fit to the data, T has seven degrees of freedom. The probability of observing a value of $T \geq 19.1285$, with seven degrees of freedom, is 0.0078. On the basis of this test, we should reject H_0.

Although the result of the chi-square test contradicts three out of the four graphical analyses, it is consistent with the percent error probability plot (Figure 1-6a). The reason for this concordance has to do with the underlying computations. Whereas T is based on the squared deviations in the cell frequencies, the percent error is based on the simple deviations of the cell probabilities. As a consequence, the two statistics differ primarily in the presence of the sign (±) of the deviation. Thus, while both Figure 1-6a and row 1 of Table 1-2 indicate large deviations in the first percentile, only Figure 1-6a demonstrates that the deviation is due to underestimation by the model.

The primary reason for the small *p*-value on the chi-square test is sample size. The individual cell chi-squares are multiples of the total sample size, n=2,541. If the same sized deviations from the model had been observed in a smaller but still respectably sized sample of n=250, the resulting chi-square statistic would have been 1.88 with P=0.9662. This illustrates the well-known maxim that no model ever fits the data. As more data are accumulated in the sample, the data will eventually depart from *any* model. The best one can hope for is a reasonable approximation. As with regression models, it is recommended that interpretation of GOF be based on careful consideration of *both* graphical summaries and formal tests of GOF hypotheses. This is the approach that is applied throughout this report.

1.5 Uncertainty in Monte Carlo Risk Assessment Models

Deterministic risk models are algebraic expressions wherein the input factors are point estimates of attributes of the population at risk (PAR) and of the risk factors themselves. Monte Carlo models employ similar algebraic expressions but their input values are random variables, that is, PAR attributes and exposures are modeled as variables with known probability distributions. Such models are called stochastic models and are further distinguished from their deterministic counterparts in that their outputs are distributions of risk rather than point estimates. Stochastic models are necessarily more complex in their mathematics and data requirements but yield estimates that are far more realistic and hence more useful than deterministic models. The reason for this, of course, is that the "real world" is

beset with uncertainty and variation. However, Monte Carlo simulation techniques do not automatically ensure that the major components of natural variation and uncertainty will be accounted for. In this section, we illustrate and discuss three types of uncertainty and their importance to risk assessment modelers.

Consider a modeler who must estimate the risk for a rural population exposed to pesticides through well water. Assume that the PAR is all the residents living within 10 miles of a large agricultural operation that has been applying pesticide A to its croplands for the past 25 years. Assume further that there are 500 wells within this area. Unfortunately, the only exposure data available to the modeler come from a sample of 16 publicly owned wells in the target area. Nonetheless, our modeler proceeds and, applying techniques outlined in this report, obtains MLEs for a series of candidate parametric distribution models and finally determines that the 16 concentrations of pesticide A best fits a lognormal with mean X and standard deviation S. After completing similar estimation and fitting procedures for the other model variables, the modeler generates a distribution of 500 risk estimates from which he determines the 95th percentile of risk with 95% confidence limits. Based upon these results, local health and agricultural officials enact regulations to curtail the use of pesticide A by the farmer. The farmer takes exception to their model results and their regulations; does he have a case?

Perhaps the most apparent problem concerns *data uncertainty*. If the data are not representative of the PAR, then even the most skillfully applied state-of-the-art modeling procedures will not yield reliable risk estimates. The definition of a "representative sample" is illusive. Kruskal and Mosteller (1971) discuss the problem at length and conclude that representativeness does not have an unambiguous definition. For our discussion, we follow the suggestions of Kendall and Buckland (1971) that a representative sample is one that is typical in respect to the characteristics of interest, however chosen. But it should be recognized that, assuming a sufficient sample size, only sampling designs in which all members of the PAR have an equal chance of selection can be guaranteed to yield representative samples.

Clearly that was not the case in our hypothetical example. For valid inference, the selected wells should be typical of the PAR in time of measurement, geographical location, construction, and any other attributes likely to affect pesticide concentration. The case in point, in which only some homogenous subset of the PAR was sampled, is typical of what often occurs in practice. Truly random samples are difficult and often prohibitively expensive to obtain. Thus, the modeler often will be forced to utilize surrogate data that generally have been collected for other purposes. Monte Carlo risk

estimates based on surrogate sample data will be biased to the degree that the sample exposure characteristics differ systematically from those of the PAR. While it is sometimes possible to employ statistical adjustments (e.g., weighting) to improve surrogate samples, in many cases it is not. U.S. EPA (1999) presents a complete discussion of diagnosis of and remedies for problems associated with the use of surrogate sample data in Monte Carlo risk assessment models.

In addition to problems associated with the sampling design, the representativeness of a sample depends on the size of the sample. In general, the more variable a PAR characteristic, the larger the minimum sample size necessary to ensure representativeness. Thus, it is unlikely that a sample as small as n=16 will be sufficient to capture the variability inherent in a PAR. Relevant variance estimates may be available from existing databases or the scientific literature; in rare cases, it may be necessary to conduct a pilot study to determine minimal sample sizes. Details of sample size determination are available in most applied sampling texts (e.g., Thompson, 1992). Samples that are too small will underestimate the PAR variance and are more likely to be biased than are larger samples.

Proper selection of exposure distribution models is the focus of this report. Given 16 candidate parametric exposure models (Section 1.6), uncertainty about the identity of the "true" PAR exposure model is a major concern. Risk distributions obtained from an incorrect exposure distribution model may be severely biased. However, properly applied estimation and GOF techniques should reduce *model uncertainty* to acceptable levels. Models with differing numbers of parameters can be compared and selected with LRTs. Selection among competing models with same number of parameters can be made on the basis of the size of the chi-square GOF p-value and the plots described in Section 1.4. However, it is possible to obtain nearly identical fits from different models. Examples of the close similarity among some models was illustrated in Section 1.1 and Figure 1-3. If the goal of the modeler is to predict the risk distribution and if the pattern and size of the observed-predicted deviations are similar among two or more competing distributions, it can be argued that it does not matter which one the modeler chooses. However, if a causal model is desired, such that the parameters represent underlying physiologic, social, and/or environmental processes, then proper discrimination among well-fitting models will be crucial. Fortunately, the vast majority of risk assessments are predictive in nature so the modeler does not need to be too concerned about very fine differences in fit among good-fitting models.

Because estimates of population parameters are based on sample data, any estimate, regardless of how it is obtained (MLE, WLS, MOM, etc.), will be subject to sampling error. Accordingly, a Monte

Carlo risk distribution estimated from an exposure model that has been correctly fit to a representative sample still will be subject to the effects of *parameter uncertainty* in the fitted exposure model. To account for these effects, it is necessary to estimate the sampling distribution of the model parameters. If ML parameters are employed, asymptotic normality can be invoked and confidence limits on the parameters can be computed as described in Section 1.3. Values of the parameters within the 95% confidence limits then can be used in a sensitivity analyses of the exposure distribution model. Alternatively, acceptable parameter values can be drawn from the multivariate normal distribution centered at the parameter MLE, with variance-covariance matrix equal to the inverse of the information matrix.

If asymptotic normality cannot be assumed either because the sample size is too small (e.g., n=16) or because MLEs were not (or could not) be obtained, bootstrap methods should be employed. The bootstrap is a versatile nonparametric method that can be used in a wide variety of situations to obtain the sampling distribution of any model parameter. For a given sample size n, some number (e.g., 1,000) of bootstrap samples, each of size n, are obtained by sampling, with replacement, from the original sample. A new estimate of the model parameter is obtained from each bootstrap sample, thereby generating a distribution of 1,000 bootstrap parameter estimates. Finally, nonparametric bias-adjusted techniques are used to compute the standard error and confidence intervals about the original parameter point estimate. Details of the bootstrap method are available in Efron and Gong (1983) or in a more user friendly format in Dixon (1993). Bootstrapping programs can be implemented easily with commercial statistical software such as SAS or SPLUS.

1.6 Summary of a System for Fitting Exposure Factor Distributions

The system of options includes components for models, estimation, assessment of fit, and uncertainty. The methods of estimation, testing of GOF, and uncertainty that we regard as most useful are printed in boldface.

1.6.1 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 nonconsuming or nonexposed individuals. Appendix D contains a table of relevant functions for

calculation of probabilities and moments (e.g., means and variances) of models in the generalized F hierarchy. To analyze a large number of EFH datasets, it may be possible and advisable to use a smaller set of models. The number of free or adjustable parameters for each model is given in parentheses, below.

- Models with a point mass at zero:
  - Generalized F (5)
  - Generalized gamma (4)
  - Burr (4)
  - Gamma, lognormal, Weibull, log-logistic (3)
  - Exponential (2)

- Models without a point mass at zero:
  - Generalized F (4)
  - Generalized gamma (3)
  - Burr (3)
  - Gamma, lognormal, Weibull, log-logistic (2)
  - Exponential (1)

1.6.2 Methods of Estimation of Model Parameters

- **Maximum likelihood estimation**
- Minimum chi-square estimation
- Weighted least squares estimation
- Minimum distance estimation
- Method of moments estimation
- Meta-analysis
- Regression on age and other covariates

1.6.3 Methods of Assessing Statistical GOF of Probability Models

- **Probability-probability plots, quantile-quantile plots, percent error plots**
- **Likelihood ratio tests 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

1.6.4 Methods of Estimating Uncertainty in the Model Parameters

- **Asymptotic normality of parameter estimates**
- **Bootstrapping from the estimated model**
- Simulation from the normalized likelihood
- Meta-analysis to combine multiple sources or studies

1.6.5 System Output

- Recommended type of model
- Estimated distribution for model parameters

The system is discussed in more detail in Section 2. Section 2 is fairly technical and may be skimmed. Applications to drinking water, population mobility, and inhalation rates are discussed in Sections 3, 4, and 5, respectively. Section 6 discusses additional issues, such as the feasibility of applying the procedures as a production process to a large number of EFH factors.

Table 1-1. Three MLEs of Prevalence, Given Different Observed Numbers of Infections

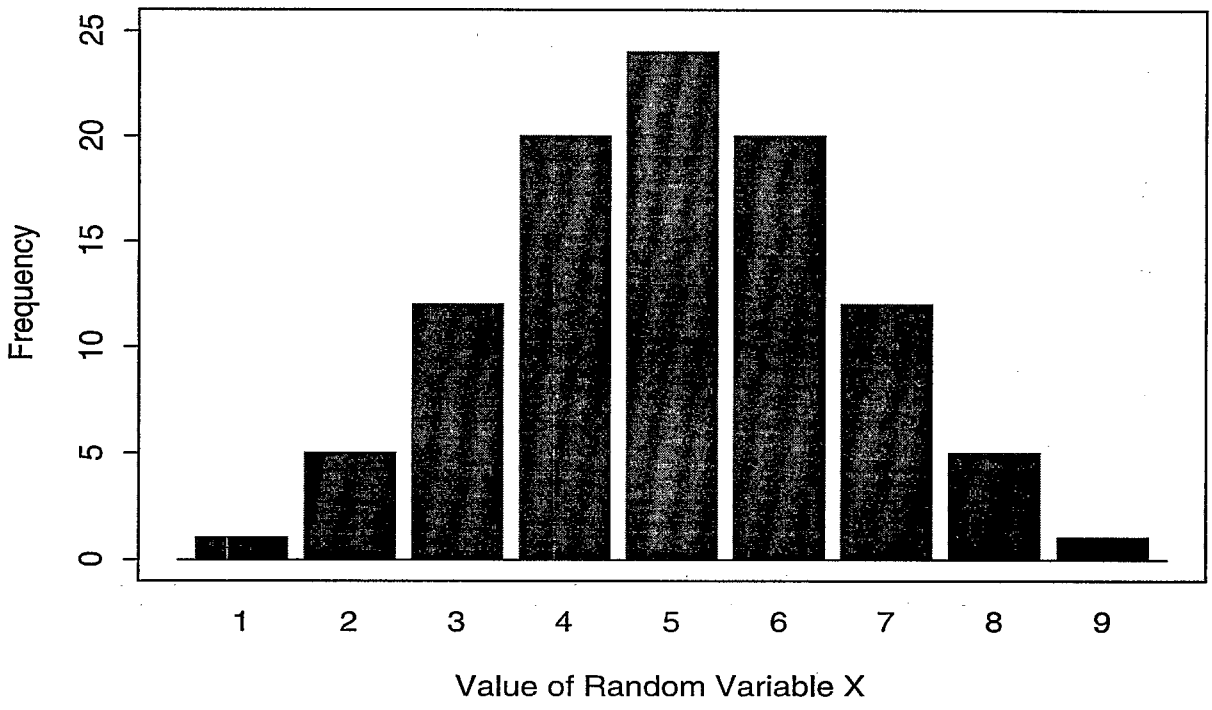
Obs. No. Infections in Sample of Nine	Likelihood of Infection			MLE of Pop. Prevalence (P)
	If P=20%	If P=50%	If P=80%	
0	0.134	0.002	0.000	0.20
1	0.302	0.018	0.000	0.20
2	0.302	0.070	0.000	0.20
3	0.176	0.164	0.003	0.20
4	0.066	0.246	0.017	0.50
5	0.017	0.246	0.066	0.50
6	0.003	0.164	0.176	0.80
7	0.000	0.070	0.302	0.80
8	0.000	0.018	0.302	0.80
9	0.000	0.002	0.134	0.80
	1.000	1.000	1.000	

Table 1-2. Computation of Chi-Square GOF for Tap Water Consumption by Persons 65 Years or Older; the Hypothesized Probability Model Is a Gamma Distribution (MLE[SCALE]=4.99731, MLE[SHAPE]=0.04365)

% Tile	Tap Water Consump.	Observ. Prob.	Pred. Prob.	Obs. N	Gamma Exp. N	Cell Chi-Sq.
1	4.5	0.01	0.00420	25.41	10.67	8.5479
5	8.7	0.04	0.04777	101.64	121.38	3.8348
10	10.9	0.05	0.05651	127.05	143.60	2.1561
25	15.0	0.15	0.15457	381.15	392.76	0.3535
50	20.3	0.25	0.23312	635.25	592.35	2.8972
75	27.1	0.25	0.24666	635.25	626.76	0.1134
90	34.7	0.15	0.15509	381.15	394.08	0.4383
95	40.0	0.05	0.05283	127.05	134.25	0.4084
99	51.3	0.04	0.04045	101.64	102.79	0.0129
100		0.01	0.00880	25.41	22.36	0.3659
		1.00	1.00000	2541.00	2541.00	19.1285

Figure 1-1. Histograms and PDFs

(a) Frequency Distribution of a Random Variable



(b) Relative Frequency Histogram and Plot of Normal PDF

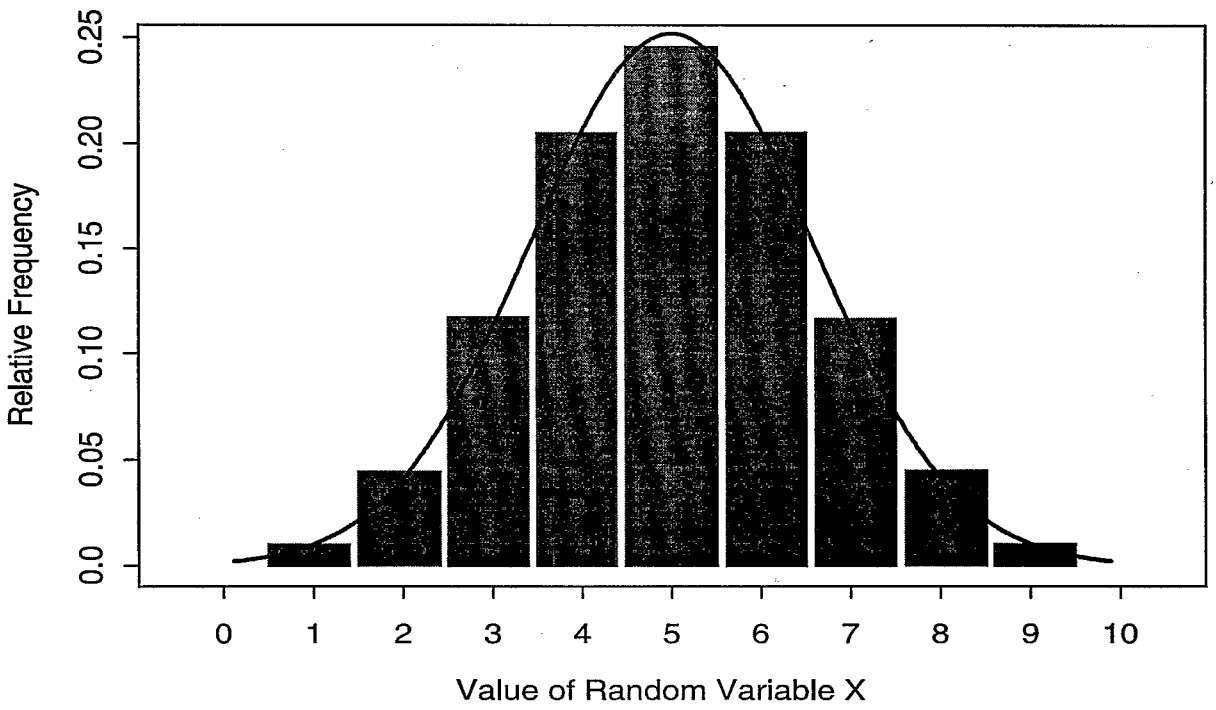
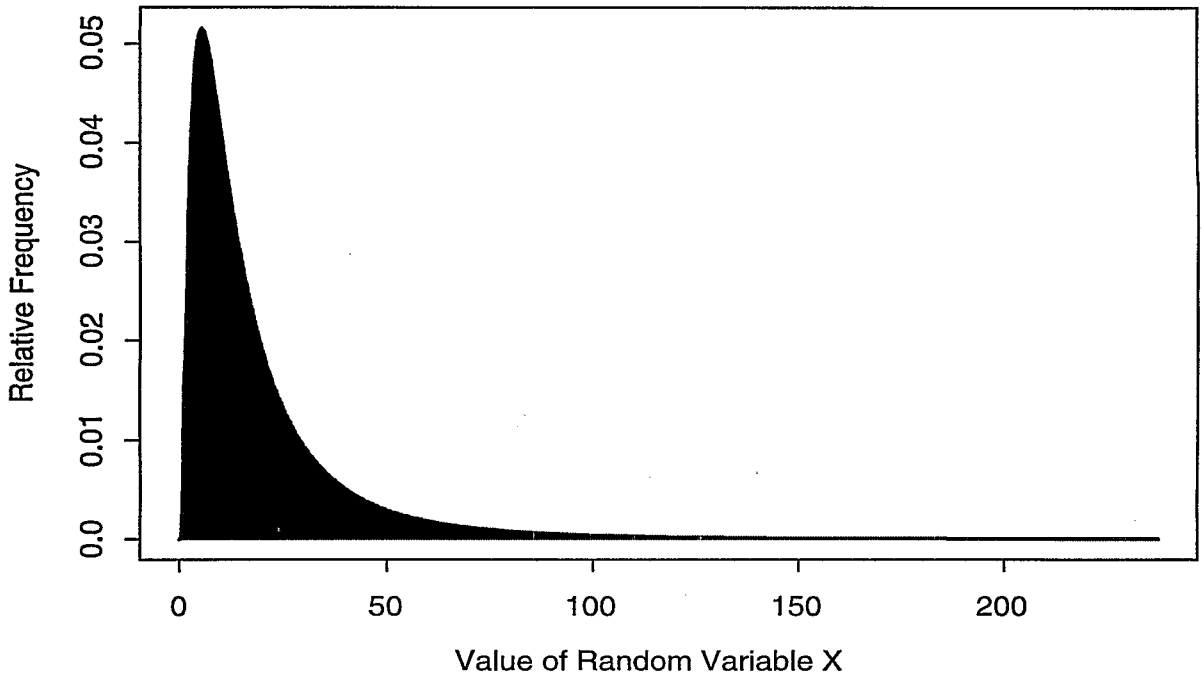


Figure 1-2. PDFs and CDFs

(a) Lognormal PDF (Mean=20, Std. Dev.=24)



(b) Lognormal CDF (Mean=20, Std. Dev.=24)

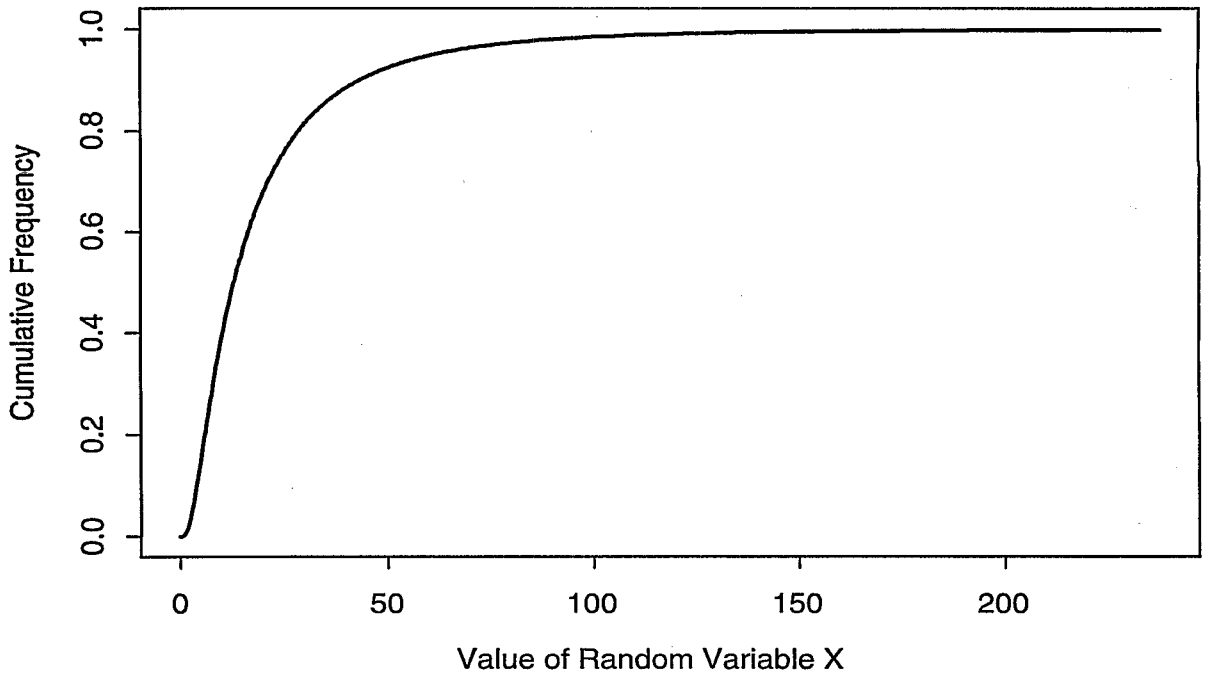


Figure 1-3. Examples of Parametric PDFs

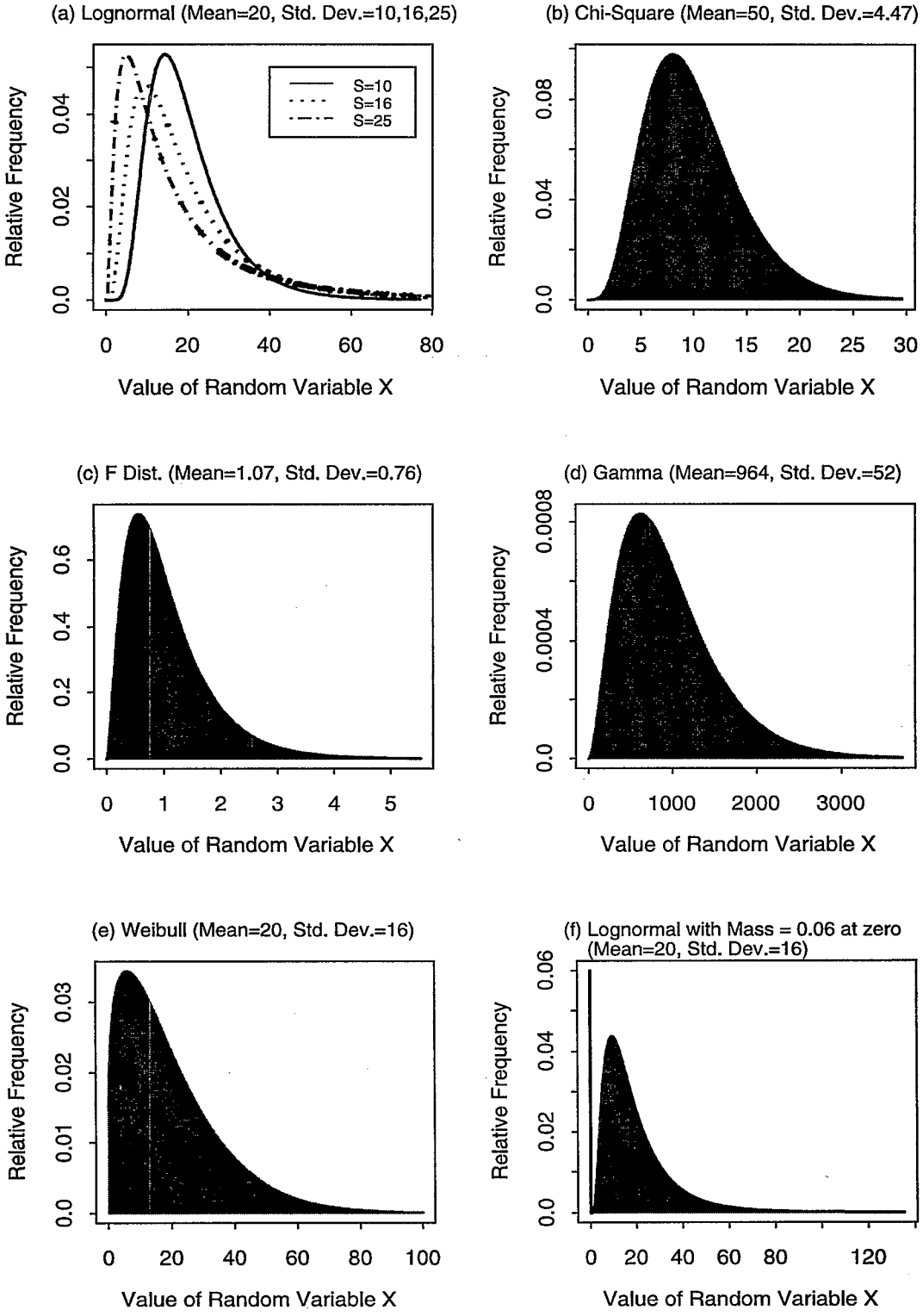


Figure 1-4. Demonstration of MLE

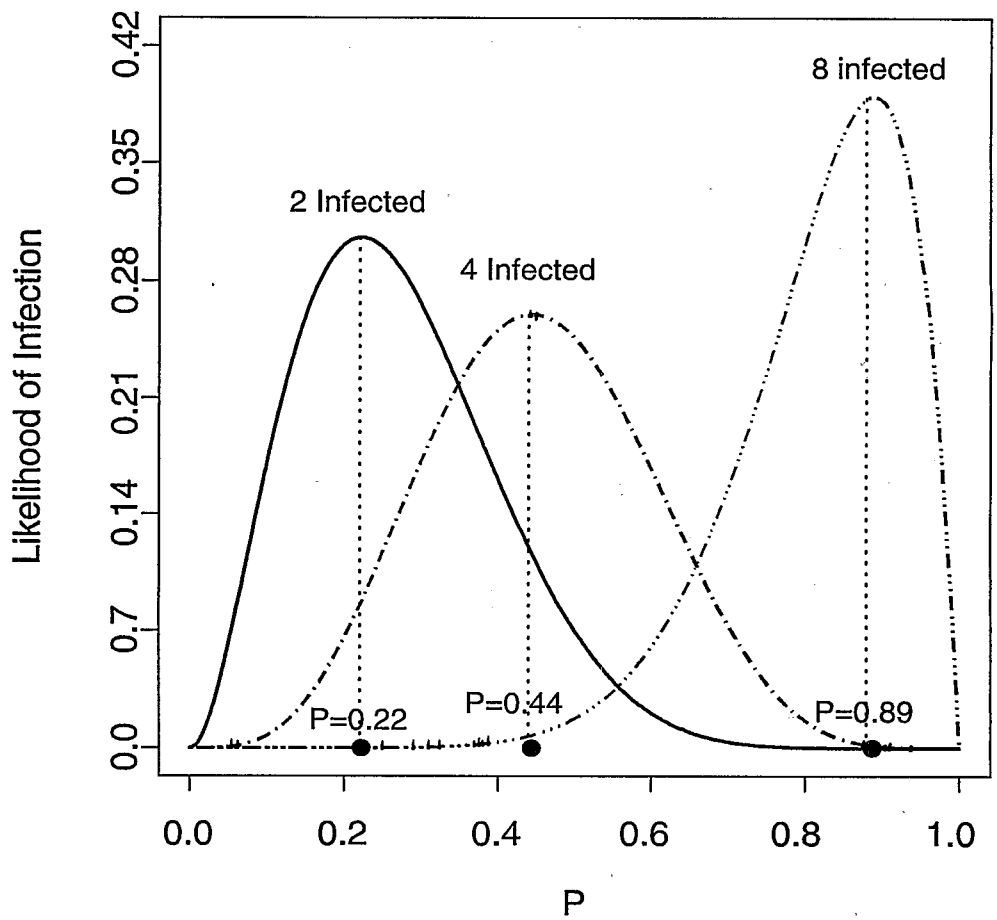
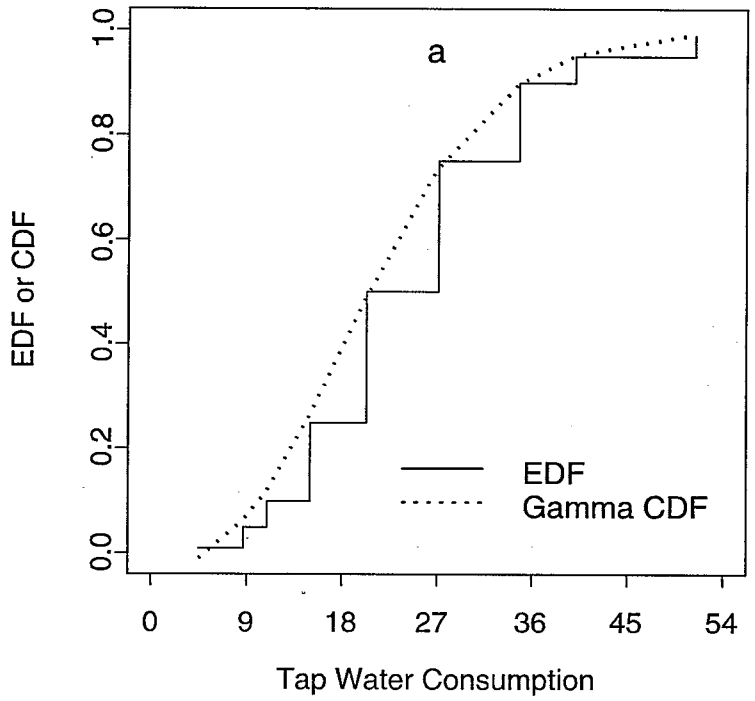


Figure 1-5. Tap Water Gamma GOF Plots: Adults 65 Years Old and Older



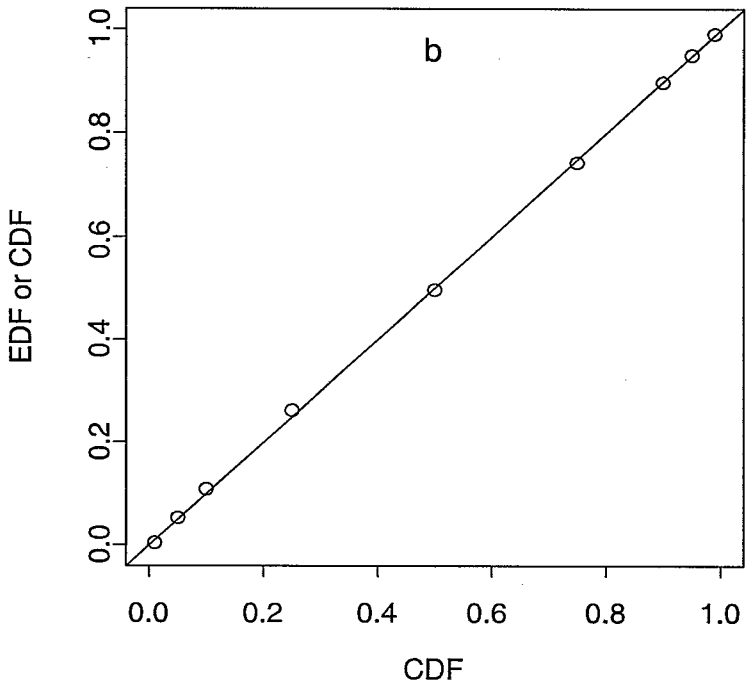
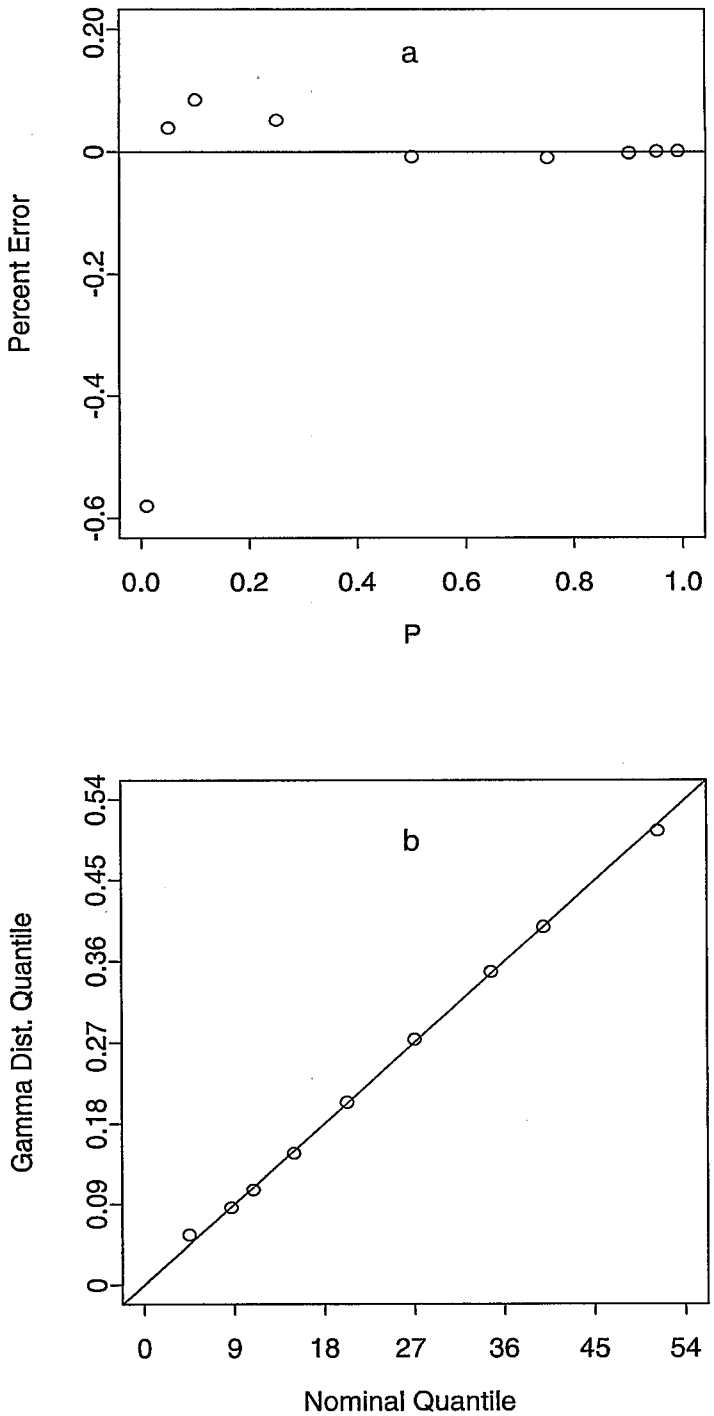


Figure 1-6. Tap Water Gamma P-P Plot and Q-Q Plot: Adults 65 Years Old and Older



2

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. Each model contains the models below it as special cases (except that the generalized gamma does not include the log-logistic). 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. For example, the parameter vector can be chosen to minimize the Anderson-Darling or Cramer-von Mises distance between an empirical and theoretical distribution function. Such MDEs are reputedly robust to model violations, but their distribution theory is less accessible than that for MLEs and MCS estimates.

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.

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 $(\hat{P}\text{-}P)/P$ 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

- Asymptotic normality of parameter estimates
- Bootstrapping
- Simulation from the normalized likelihood
- 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.

Although uncertainty in risk assessment has been discussed extensively, a consensus for its treatment has not yet emerged. A simple theoretical model is not expected to capture a complex real-world situation exactly. 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. 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 only 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 of 100 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 95^{th} 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

3

Analysis of Tap Water Data

3.1 Methods

Here and throughout this report, the statistical summaries from the Exposure Factors Handbook (EFH) are analyzed. No attempt was made to obtain raw data from investigators.

The key studies identified in the EFH are Canadian Ministry of National Health and Welfare (1981) and Ershow and Cantor (1989). Since the first dataset is Canadian, is older, and involves a much smaller sample size, it was decided to base the analysis only on the second dataset. Specifically, the focus was on the six age groups at the bottom part of Table 3-7 in the EFH, which has age categories for infants (age <1), children (ages 1-10), teens (ages 11-19), younger adults (ages 20-64), and older adults (ages 65+), as well as all ages. The EFH Table 3-7 data summaries analyzed here consist of nine estimated percentiles for total daily tap water intake in dL/kg/day. (EFH Table 3-7 units are mL/kg/day; these were rescaled to dL/kg/day to obtain better convergence properties for numerical optimization routines.) The tabulated percentiles from EFH Table 3-7 are reproduced in this report in Table 3-5, columns labeled "X_p = Data Qtile" and "Nom p" (for "Data Quantile" and "Nominal p"). These percentiles correspond to probabilities of 0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95, and 0.99. That is, Xp is the tap water consumption value such that 100p% of the population consumes Xp or less daily, or the tap water consumption value such that the cumulative distribution function (CDF) value is p at Xp, F(Xp)=p. For example, referring to Table 3-5, the 25th percentile for adults of ages 20-64 is 0.124, so that approximately one-fourth of adults between ages 20 and 64 consume 12.4 mL/kg/day or less of tap water. Only six percentiles are shown for infants because the 1st, 5th, and 10th percentiles are all zero for infants. This motivates the inclusion of a point mass at zero in probability models as discussed in Section 1.

The 12 models of the generalized F hierarchy were fit to each of the six tap water datasets from the bottom of EFH Table 3-7 using three different estimation criteria—maximum likelihood estimation (MLE), minimum chi-square (MCS) estimation, and weighted least squares (WLS). The Pearson chi-

square tests and likelihood ratio tests (LRTs) of goodness-of-fit (GOF) were used. These models, estimation criteria, and GOF tests are discussed in Section 2.

Because the sample size was quite large, the asymptotic normality approach was used to obtain parameter uncertainty distributions. The two-step simulation process was applied 10,000 times to obtain simulated distributions of drinking water values for each age group. Quantiles corresponding to the same nine nominal probability values (0.01, 0.05, . . ., 0.99) were determined from the simulated drinking water distributions. Models were fit to these simulated quantiles using the same MLE technique that was applied to the empirical percentiles. Model-based averages, standard deviations, and quantiles were estimated from the simulated data and compared with those estimated from the percentile data.

3.2 Results

The three methods of estimation (MLE, MCS, and WLS) and two methods of testing fit (chi-square and LRT) led to essentially the same conclusions regarding fit of the different models. Therefore, only results from the chi-square GOF test based on the MLE are shown.

Values of the chi-square statistic and associated *p*-values for chi-square GOF tests are provided in Tables 3-1a and 3-1b. In each case, the null hypothesis tested is that the data arose from the given type of model. A low p-value casts doubt on the null hypothesis. Clearly, the only model that appears to fit most of the datasets is the five-parameter generalized F distribution with a point mass at zero, referred to as GenF5. This point is illustrated graphically via probability-probability (P-P), quantile-quantile (Q-Q), and percent error plots in Figures 3-1 and 3-2 (figures are at the end of Section 3).

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. For the tap water data, the nominal probabilities are 0.01, 0.05, 0.10, etc. Q-Q plots show the model-based quantile estimates on the vertical axis versus empirical quantiles (Xp values) on the horizontal axis. For the tap water data for adults between ages 20 and 64, the empirical quantiles corresponding to nominal probabilities of 0.01 and 0.05 are 0.022 and 0.059. In addition to P-P and Q-Q plots, Figures 3-1 and 3-2 also show the corresponding percent error plots, that is, plots of $(\hat{P}\text{-}P)/P$ versus P and plots of $(\hat{Q}\text{-}Q)/Q$ versus Q. As explained in Section 2.3.1, the region of interest in P-P and Q-Q plots is near the main diagonal, and percent error plots are more informative because they transform and magnify this region. The term

percent error is used loosely, because the plotted quantities are error fractions as opposed to percents (e.g., 1.5 and -1.5 are plotted to represent 150% and -150%).

If possible, it is desirable to use one of the standard two-parameter models (gamma, lognormal, Weibull), unless there is strong evidence that a model with more parameters is required. Results of this analysis have shown, in fact, that the five-parameter generalized F distribution with a point mass at zero provides considerably better fit to the tap water data than any of these two-parameter models. However, risk assessors might still prefer to use the two-parameter models, on grounds of simplicity and familiarity.

According to Table 3-1a, the gamma model provides the best fit (smallest chi-square) of the two-parameter models to the data for each of the five individual age groups. For the group with all ages pooled, the log-logistic and gamma are the best and second-best fitting two-parameter models.

Table 3-2 summarizes several additional aspects of interest for the tap water populations. Within each age group, the first row (SOURCE=data) is basically a data summary. Within the top row, the columns labeled N, MEAN, and SDEV contain the sample size, the sample mean, and the sample standard deviation. Within the top row, the columns labeled P01, P05, . . ., P99 contain the nominal probabilities 0.01, 0.05, . . ., 0.99. The values in the top row for MEAN, SDEV, and the nine nominal probabilities can be thought of as 11 targets that the models are trying to hit.

In Table 3-2, the other five rows (second through sixth rows) within each age group contain results from fitting four models, including gamma, lognormal, and Weibull, using selected estimation criteria. The model and estimation criteria are indicated by the variable SOURCE. For instance, SOURCE=gammle indicates the two-parameter gamma model fit using MLE. The model gf5 is the five-parameter generalized F with a point mass at zero. The infants group does not contain results from the five-parameter generalized F because the model selected had infinite variance. For the gamma and Weibull models, there was little difference between the three estimation criteria, and the MLE performed best overall. For the lognormal model, results from the WLS estimation criterion are shown in addition to the MLE. These will be contrasted below.

The last two columns contain summary GOF measures. CHIDF is the value of the chi-square statistic divided by its degrees of freedom. The methods are ordered with respect to this CHIDF

measure. CHIDF is more comparable across cases involving different degrees of freedom than is the chi-square statistic. PGOF is the *p*-value for model GOF based on the chi-square test. Low-values of PGOF, such as PGOF<0.05, cast doubt on the null hypothesis that the given type of model is correct.

Note that MLE performed much worse for the lognormal model than the WLS method of estimation, as determined by CHIDF and PGOF measures.

If a two-parameter model must be used for tap water consumption, then the gamma model with parameters estimated by maximum likelihood is recommended. The five-parameter generalized F distribution could be used for sensitivity analyses.

The age effect seems sufficiently strong to justify the use of separate age groups in risk assessment. Note, however, that the lognormal model with parameters estimated by WLS provides the best fit among the two-parameter models, as determined by CHIDF, when all age groups are pooled.

3.3 Uncertainty Analysis

Table 3-3 contains information on the uncertainty distribution parameters of the best fitting two-parameter distributions, namely, the gamma distributions. The parameter estimates log α and log β are the MLEs of the natural logs of the usual gamma parameters α and β. The variables SELα and SELβ are the standard errors of these estimates, and CORR is the estimated correlation between the parameter estimates. To generate values for the gamma parameters, first values for the logarithms of α and β are generated by sampling from a bivariate normal distribution with mean parameters log α and log β, with standard deviations SELα and SELβ, and correlation CORR. The generated values of log α and log β are then exponentiated to obtain values for α and β.

Because the underlying sample sizes are quite large, these parameter uncertainty distributions based on asymptotic normality are probably adequate. Comparisons with bootstrap and likelihood methods via simulation studies could shed light on this issue.

Tables 3-4 and 3-5 contain results from the original data analysis and from the two-step simulation process based on asymptotic normality, using the bivariate normal distributions summarized in Table 3-3 to represent distributions of parameter uncertainty. For each age group, 10,000 drinking

water values were generated by first drawing a parameter pair ($\log\alpha$ and $\log\beta$) from the bivariate normal distribution of Table 3-3, then generating a drinking water value from the selected gamma distribution. Next, the nine nonparametric quantiles were estimated for each age group from the samples of size 10,000. Gamma distributions were fit to these quantiles using the same maximum likelihood method that was applied in the original analysis described in Section 3.2.

Tables 3-4 and 3-5 show that the results of the two-step process are very similar to the original fitted gamma distributions. Table 3-4 contains data means and standard deviations as well as MLEs of the means and standard deviations from the original analysis of the data (MLE Mean and MLE Sdev) and from the analysis of the simulated data from the two-step process (MLE2 Mean and MLE2 Sdev). In all cases, except infants, MLE and MLE2 agree to within 0.002.

Table 3-5 contains several estimates of quantiles as well as two estimates of the CDF evaluated at the pth quantile, $F(x_p)$. As before, X_p denotes the original empirical pth quantile from EFH Table 3-7. (In theory, if x_p were the true quantile, then $F(x_p)=p$.) The other quantile estimates are the MLE from the original data analysis (MLE Qtile), the nonparametric quantiles from the simulated data (two-step Empl Qtile) that incorporate parameter uncertainty, and the MLE for the simulated data (MLE2 Qtile). The last two columns contain MLEs of $F(x_p)$ from the original data analysis and from the simulated data. Except for the teens group, these MLES of $F(x_p)$ always agree to within 0.004.

In general, the values of the MLEs of quantities estimated from the original analysis of the raw data and from the simulated data reflecting parameter uncertainty are very close. Presumably, this is a consequence of the large sample sizes underlying the raw data.

3.4 Conclusions

The tap water data from EFH Table 3-7 force a difficult question: How good does the fit need to be? Among two-parameter models, the gamma distribution fits best. The two-parameter gamma model may fit well enough for most purposes. However, it is also true that this model fails to pass the chi-square GOF test, while the five-parameter generalized F distribution passes at the 0.05 level in four of six cases.

If the situation warrants a more sophisticated model, the generalized F may be used. However, the uncertainty analysis for the five-parameter model could be complicated. The five-parameter model

entails very highly correlated parameters. Contours of the likelihood in five-space might be highly nonelliptical. One would not be comfortable with an uncertainty analysis for the five-parameter model based on asymptotic normality without investigating its behavior by additional simulation studies. Another possibility worth investigating would be uncertainty analysis for the five-parameter model based on bootstrapping. According to Efron and Tibshirani (1993), the parametric bootstrap will automatically endow the right shape to the simulated distribution for the parameters, although bias correction may be needed if the simulated distribution is not centered at the original parameter estimates.

The distributions presented in this section for tap water intake were derived based on data of Ershow and Cantor (1989). These data were obtained from the U.S. Department of Agriculture 1977-78 Nationwide Food Consumption Survey (USDA, 1984). The main limitations of the data are that they are old and do not reflect the expected increase in the consumption of bottled water and soft drinks. The survey has, however, a large sample size (26,466 individuals), and it is a representative sample of the U.S. population with respect to age distribution, sex, racial composition, and regions. Therefore, these distributions are applicable to cases where the national tap water consumption is the factor of interest or it can reasonably be assumed that the population of interest will have consumption rates similar to the national U.S. population.

Table 3-1a. Chi-Square GOF Statistics for 12 Age-Specific Models, Fit to Tap Water Data, Based on Maximum Likelihood Method of Parameter Estimation[a]

Age Group	CHI Gam2	CHI Log2	CHI Tic2	CHI Wei2	CHI Ggam3	CHI GenF4	CHI Gam3	CHI Log3	CHI Tic3	CHI Wei3	CHI Ggam4	CHI GenF5
Infants (<1)	19.8	26.6	39.4	20.6	18.1	10.6	19.8	13.7	10.8	20.6	18.1	8.10
Children (1-10)	84.5	315	295	198	84.7	40.3	46.6	129	195	198	27.5	15.2
Teens (11-19)	89.5	606	557	125	81.4	38.4	23.4	286	377	110	23.1	7.88
Adults 1 (20-64)	144	734	719	319	139	38.8	42.8	354	491	319	42.1	3.96
Adults 2 (65+)	19.2	83.3	101	107	20.2	9.72	5.08	30.1	73.0	107	2.16	1.24
All	847	1180	597	1807	780	154	550	473	251	1807	313	6.36

[a]Prefix indicates model type: Gam = gamma, Log = lognormal, Tic = log-logistic, Wei = Weibull, Ggam = generalized gamma, GenF = generalized F.

Numeric model suffix indicates number of free or adjustable parameters.

Degrees of freedom for X^2GOF=number of quantile categories – number of model parameters.

Table 3-1b. P-Values for Chi-Square GOF Tests of 12 Age-Specific Models, Tap Water Data[a]

Age Group	PGOF Gam2	PGOF Log2	PGOF Tic2	PGOF Wei2	PGOF Ggam3	PGOF GenF4	PGOF Gam3	PGOF Log3	PGOF Tic3	PGOF Wei3	PGOF Ggam4	PGOF GenF5
Infants (<1)	0.001	0.000	0.000	0.000	0.000	0.005	0.000	0.003	0.013	0.000	0.000	0.013
Children (1-10)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004
Teens (11-19)	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.096
Adults 1 (20-64)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.412
Adults 2 (65+)	0.008	0.000	0.000	0.000	0.003	0.084	0.533	0.000	0.000	0.000	0.827	0.871
All	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.174

[a]Prefix indicates model type: Gam = gamma, Log = lognormal, Tic = log-logistic, Wei = Weibull, Ggam = generalized gamma, GenF = generalized F.

Model suffix indicates number of free or adjustable parameters.

Table 3-2. Results of Statistical Modeling of Tap Water Data Using Five-Parameter Generalized F and Two-Parameter Gamma, Lognormal, and Weibull Models[a]

SOURCE	N	P01	P05	P10	P25	P50	P75	P90	P95	P99	MEAN	SDEV	CHIDF	PGOF
INFANTS (Age <1)														
data	403	.010	.050	.100	.250	.500	.750	.900	.950	.990	.435	.425		
gammle					.252	.526	.702	.908	.951	.996	.448	.410	4.945	.0006
weimle					.260	.526	.699	.906	.950	.996	.447	.412	5.145	.0004
logmle					.227	.561	.735	.903	.937	.984	.470	.548	6.660	.0000
logwls					.216	.559	.738	.908	.942	.986	.462	.512	6.974	.0000
CHILDREN (Ages 1-10)														
data	5605	.010	.050	.100	.250	.500	.750	.900	.950	.990	.355	.229		
gf5mle		.010	.047	.106	.250	.495	.752	.900	.952	.989	.356	.234	3.792	.0044
gammle		.004	.052	.118	.263	.492	.738	.895	.953	.993	.355	.224	12.07	.0000
logwls		.000	.024	.091	.266	.529	.765	.895	.943	.984	.356	.250	27.18	.0000
weimle		.011	.070	.134	.264	.474	.721	.894	.959	.997	.355	.218	28.34	.0000
logmle		.000	.036	.113	.288	.532	.750	.878	.929	.977	.366	.286	45.07	.0000
TEENS (Ages 11-19)														
data	5801	.010	.050	.100	.250	.500	.750	.900	.950	.990	.182	.108		
gf5mle		.010	.048	.103	.253	.498	.747	.902	.953	.989	.182	.110	1.969	.0962
gammle		.002	.046	.110	.274	.511	.740	.891	.947	.989	.182	.111	12.79	.0000
weimle		.006	.061	.122	.267	.487	.725	.895	.957	.995	.182	.106	17.86	.0000
logwls		.000	.017	.076	.270	.544	.768	.896	.942	.981	.182	.119	45.35	.0000
logmle		.000	.032	.108	.303	.548	.747	.871	.920	.968	.189	.144	86.56	.0000
ADULTS 1 (Ages 20-64)														
data	11731	.010	.050	.100	.250	.500	.750	.900	.950	.990	.199	.108		
gf5mle		.010	.051	.098	.251	.501	.748	.901	.951	.990	.199	.110	0.990	.4116
gammle		.003	.049	.105	.270	.510	.738	.891	.947	.992	.199	.109	20.50	.0000
weimle		.010	.069	.122	.267	.484	.719	.893	.957	.997	.199	.105	45.54	.0000
logwls		.000	.024	.079	.273	.542	.762	.893	.941	.984	.199	.116	69.20	.0000
logmle		.000	.037	.100	.295	.543	.747	.875	.925	.976	.203	.132	104.9	.0000

Table 3-2. Results of Statistical Modeling of Tap Water Data Using Five Parameter Generalized F and Two-Parameter Gamma, Lognormal, and Weibull Models[a] (continued)

SOURCE	N	P01	P05	P10	P25	P50	P75	P90	P95	P99	MEAN	SDEV	ADJCHI	PGOF
							ADULTS 2 (Ages 65+)							
data	2541	.010	.050	.100	.250	.500	.750	.900	.950	.990	.218	.098		
logwls		.000	.032	.090	.267	.524	.762	.898	.944	.984	.218	.102	0.237	.0000
gf5mle		.010	.049	.101	.253	.496	.750	.902	.951	.989	.218	.098	0.310	.8715
logmle		.001	.041	.104	.280	.525	.751	.886	.934	.979	.220	.109	1.900	.0000
gammle		.004	.052	.109	.263	.497	.742	.898	.950	.991	.218	.098	2.746	.0075
weimle		.017	.079	.132	.262	.467	.717	.898	.960	.997	.218	.097	15.270	.0000
							ALL							
data	26081	.010	.050	.100	.250	.500	.750	.900	.950	.990	.226	.154		
gf5mle		.010	.050	.099	.252	.499	.749	.902	.951	.989	.227	.168	1.589	.1740
logwls		.000	.029	.091	.278	.524	.744	.890	.945	.991	.226	.154	113.400	.0000
gammle		.003	.058	.118	.274	.491	.718	.890	.955	.997	.225	.138	121.000	.0000
logmle		.000	.041	.112	.299	.529	.734	.875	.932	.986	.231	.173	168.600	.0000
weimle		.011	.081	.141	.281	.476	.698	.885	.958	.999	.225	.137	258.100	.0000

[a]Within each age group, the first row (SOURCE=data) is basically a data summary. Within the top row, the columns labeled N, MEAN, and SDEV contain the sample size, the sample mean, and the sample standard deviation. Within the top row, the columns labeled P01, P05, . . . , P99 contain the nominal probabilities 0.01, 0.05, . . . , 0.99. The values in the top row for MEAN, SDEV, and the nine nominal probabilities can be thought of as 11 targets that the models are trying to hit. The other five rows (second through sixth rows) within each age group contain results from fitting four models using selected estimation criteria. The model and estimation criterion are indicated by the variable SOURCE: gf5mle denotes the five-parameter generalized F distribution with a point mass at zero fit by maximum likelihood; gammle, logmle, weimle denote the two-parameter gamma, lognormal, and Weibull distributions fit by MLE; and logwls denotes the lognormal distribution fit by WLS. The last two columns contain summary GOF measures. CHIDF is the value of the chi-square statistic divided by its degrees of freedom. CHIDF is more comparable across cases involving different degrees of freedom than is the chi-square statistic. PGOF is the p-value for model GOF based on the chi-square test. Low-values of PGOF, such as PGOF <0.05, cast doubt on the null hypothesis that the given type of model is correct. Results for the generalized F distribution are not shown for infants because the estimated model had infinite variance.

Table 3-3. Uncertainty Distribution of Gamma Parameters Estimated from Tap Water Data[a]

Age Group	log (α)	log (β)	Std. Err. Log (α)	Std. Err. Log (β)	CORR (α,β)
Infants (<1)	0.1744	-0.9767	0.1738	0.2005	-0.8663
Children (1-10)	0.9221	-1.9585	0.0684	0.0757	-0.9087
Teens (11-19)	0.9889	-2.6920	0.0980	0.1077	-0.9150
Adults 1 (20-64)	1.2067	-2.8214	0.0782	0.0843	-0.9310
Adults 2 (65+)	1.6089	-3.1316	0.0555	0.0584	-0.9533
All	0.9715	-2.4653	0.1167	0.1287	-0.9143

[a]Log (α) and log (β) are MLEs of the natural logs of the gamma parameters α and β. CORR(α,β) is the estimated correlation between log (α) and log (β).

Table 3-4. Results of Two-Step Simulation Process to Incorporate Uncertainty Into Drinking Water Distributions Using Asymptotic Normality[a]

Age Group	Data Mean	MLE Mean	MLE2 Mean	Data Sdev	MLE Sdev	MLE2 Sdev
Infants (<1)	.435	.448	.451	.425	.411	.417
Children (1-10)	.355	.355	.356	.229	.224	.225
Teens (11-19)	.182	.182	.184	.108	.111	.112
Adults1 (20-64)	.199	.199	.200	.108	.109	.109
Adults2 (65+)	.218	.218	.218	.098	.098	.099
All	.226	.225	.224	.154	.138	.138

[a]MLE Mean and Sdev are MLEs of the two-parameter gamma mean and standard deviation from the original analysis.

MLE2 Mean and MLE2 Sdev are the result of the following process: generate 10,000 (α,β) pairs using the distribution of Table 3-3; for each pair, generate a drinking water value from the specified gamma distribution; calculate the nine quantiles for the resulting 10,000 drinking water values; fit a gamma distribution to the quantiles using maximum likelihood, and determine its mean and standard deviation.

Table 3-5. Uncertainty Analysis Based on Asymptotic Normality Using Two-Step Simulation Process for Two-Parameter Gamma Distributions

Age Group	X_p = Data Qtile	MLE Qtile	Empl Qtile	MLE2 Qtile	Nom p	MLE $F(x_p)$	MLE2 $F(x_p)$
Infants (<1)	.153	.152	.151	.151	.25	.252	.254
Infants	.353	.331	.332	.331	.50	.525	.525
Infants	.547	.620	.622	.624	.75	.702	.699
Infants	1.02	.989	.996	.999	.90	.908	.905
Infants	1.27	1.26	1.28	1.28	.95	.951	.949
Infants	2.21	1.89	1.93	1.92	.99	.996	.995
Children (1-10)	.027	.040	.038	.039	.01	.004	.004
Children	.083	.082	.081	.082	.05	.052	.052
Children	.125	.115	.114	.115	.10	.118	.118
Children	.196	.190	.190	.190	.25	.262	.262
Children	.305	.309	.310	.310	.50	.492	.491
Children	.460	.470	.476	.471	.75	.738	.737
Children	.644	.654	.654	.657	.90	.894	.893
Children	.794	.784	.780	.787	.95	.953	.952
Children	1.14	1.07	1.05	1.07	.99	.993	.993
Teens (11-19)	.012	.023	.022	.023	.01	.002	.002
Teens	.043	.045	.046	.046	.05	.045	.044
Teens	.065	.062	.063	.063	.10	.110	.106
Teens	.106	.100	.102	.102	.25	.274	.267
Teens	.163	.160	.162	.162	.50	.511	.503
Teens	.236	.240	.243	.243	.75	.740	.733
Teens	.323	.331	.335	.335	.90	.891	.887
Teens	.389	.395	.397	.399	.95	.947	.944
Teens	.526	.533	.536	.539	.99	.989	.988
Adults 1 (20-64)	.022	.033	.034	.034	.01	.003	.003
Adults 1	.059	.059	.060	.060	.05	.049	.048
Adults 1	.080	.078	.078	.079	.10	.105	.103
Adults 1	.124	.119	.120	.120	.25	.270	.268
Adults 1	.182	.179	.180	.180	.50	.510	.507
Adults 1	.253	.258	.257	.258	.75	.738	.737
Adults 1	.337	.345	.347	.345	.90	.891	.890
Adults 1	.400	.405	.401	.405	.95	.947	.947
Adults 1	.548	.534	.545	.534	.99	.992	.992
Adults 2 (65+)	.045	.056	.054	.055	.01	.004	.005
Adults 2	.087	.086	.085	.085	.05	.052	.055
Adults 2	.109	.106	.105	.105	.10	.109	.112
Adults 2	.150	.147	.146	.146	.25	.263	.267
Adults 2	.203	.204	.203	.203	.50	.497	.499
Adults 2	.271	.274	.274	.274	.75	.742	.742
Adults 2	.347	.349	.351	.350	.90	.898	.896
Adults 2	.400	.399	.399	.401	.95	.950	.949
Adults 2	.513	.506	.512	.510	.99	.991	.990
All	.017	.027	.027	.027	.01	.003	.003
All	.058	.054	.055	.054	.05	.058	.058
All	.082	.075	.076	.075	.10	.118	.119
All	.130	.123	.123	.123	.25	.274	.275
All	.194	.197	.196	.197	.50	.491	.491
All	.280	.296	.296	.296	.75	.718	.718
All	.398	.410	.406	.409	.90	.890	.890
All	.500	.489	.488	.489	.95	.955	.955
All	.798	.662	.682	.662	.99	.997	.997

Figure 3.1 Tap Water Intake P-P Plots: Children (EFH Table 3-7)

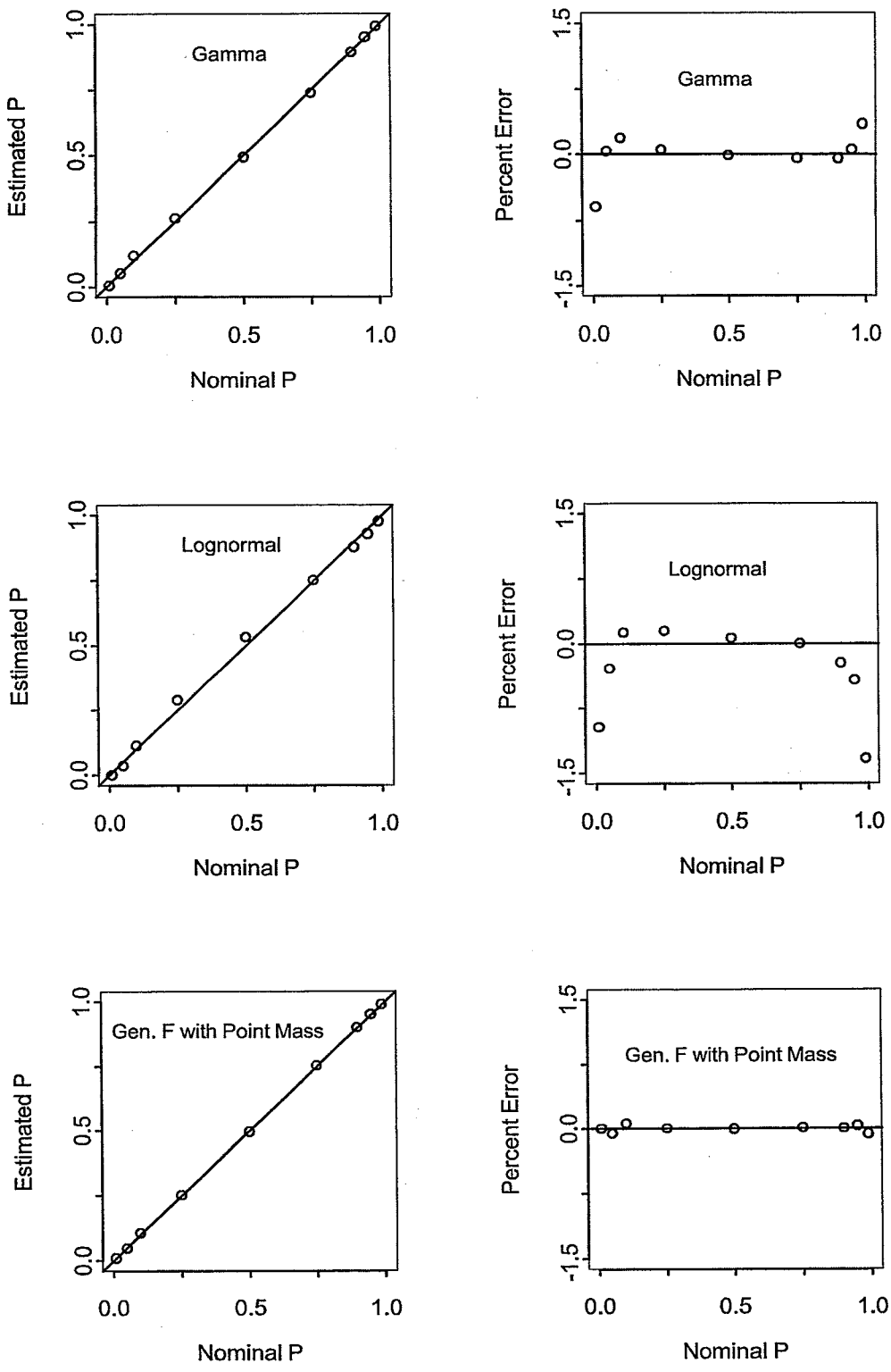
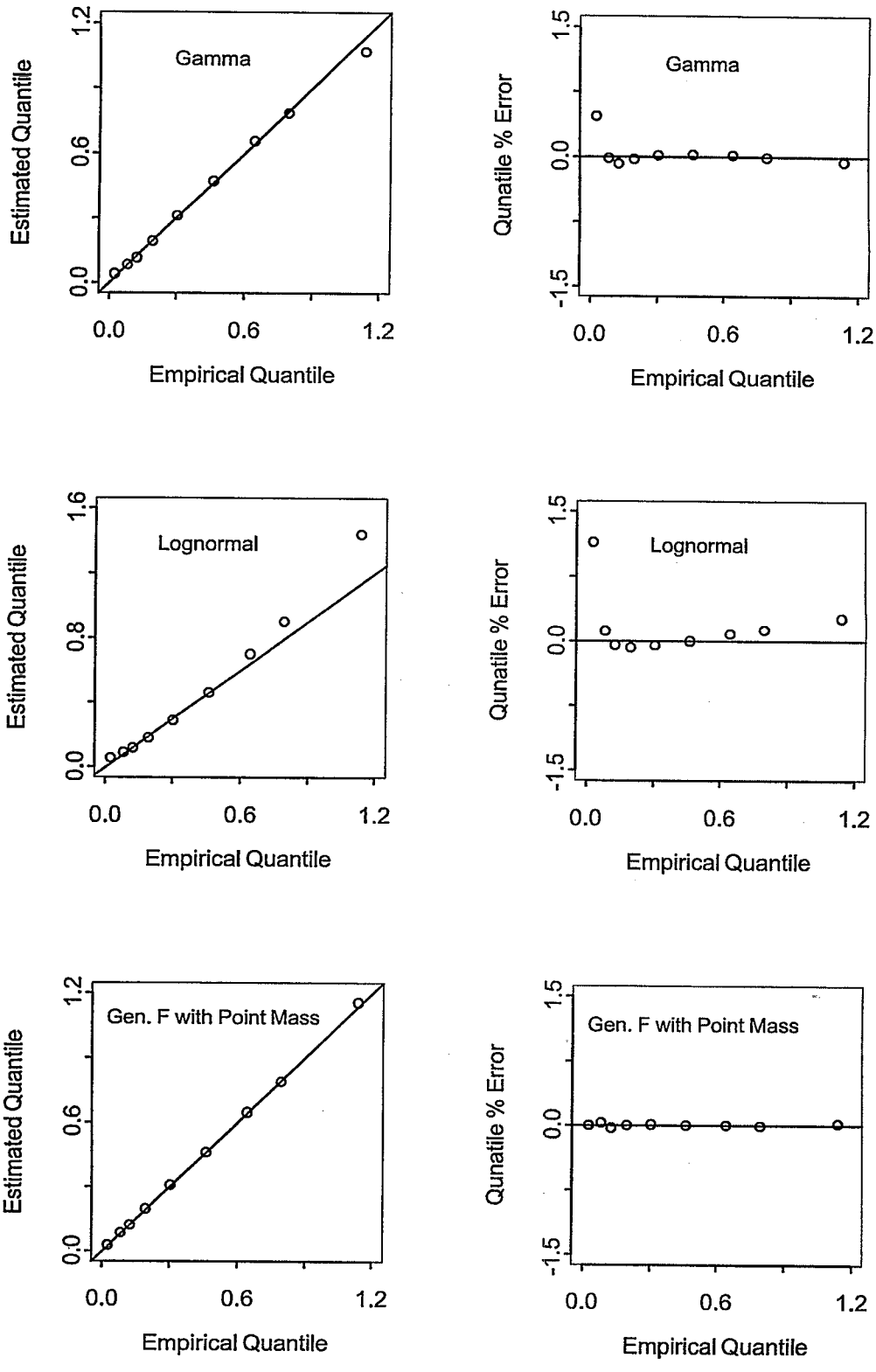


Figure 3.2 Tap Water Intake Q-Q Plots: Children (EFH Table 3-7)



4

Analysis of Population Mobility Data

4.1 Methods

4.1.1 Data

The Exposure Factors Handbook (EFH) has three key studies for population mobility. Each study uses a unique approach to define and estimate residence time. Israeli and Nelson (1992) work with current residence time (time since moving into the current residence) and total residence time (time between moving into and out of a residence). Current residence time does not seem to be directly relevant to risk assessment because it is censored; that is, the unobserved residence time is ignored. Total residence time is more relevant, but Israeli and Nelson (1992) apparently estimate it in a way that allows frequent movers to contribute more times than infrequent movers. The result is a residence time distribution that tends to be much shorter than those from the other two key studies; that is, median is 1.4 years versus a median of 9 years for each of the other two key studies.

The second key study is based on a national survey by the U. S. Bureau of the Census (1993) of 55,000 housing units that yielded 93,147 residence times. Residents were asked about time lived at current and past residences.

Johnson and Capel (1992) used a simulation model to estimate the distribution of residential occupancy periods based on a methodology described in the EFH. Occupancy period is the time between a person moving into a residence and moving out or dying. Census data were used for the dynamics of mobility. Data from the National Center for Health Statistics were used for mortality.

Table 4-1 contains estimates of selected percentiles from the three key studies. For Israeli and Nelson (1992), total residence time is used.

The residence time distributions for the second and third key studies are fairly similar at the 25th, 50th, and 75th percentiles, but the distribution of Johnson and Capel (1992) has a shorter right tail than the Census Bureau (1993) distribution. Times from Israeli and Nelson (1992) tend to be much shorter.

The first of two relevant studies (National Association of Realtors, 1993) estimated an average occupancy period of 7.1 years for homeowners. However, the response rate was only 12%. The second relevant study (Lehman, 1994) estimated average residence times as 14.3, 13.4, and 12 years for 1991, 1992, and 1993, respectively. Apparently, residence times are decreasing. The 12-year average is similar to the estimate of Johnson and Capel (1992).

Based on discussions and comparisons of the studies, Johnson and Capel (1992) seem to provide the most representative summary for EPA risk assessment purposes. They are the only source of age-specific distributions, and age is clearly a relevant factor. The analysis of population mobility data will therefore focus on the age-specific distributions of EFH Table 14-159, taken from the simulation study of Johnson and Capel (1992).

4.1.2 Statistical Methods

Models were fit to the 30 different age groups of EFH Table 14-159, which includes simulated averages and six percentiles for each group. The data of Johnson and Capel (1992) from EFH Table 14-159 are shown in Table 4-2. The simulation sample size was 0.5 million, or about 17,000 per age group. However, because their data came entirely from Monte Carlo simulations, it did not seem appropriate to treat them as if they had come from a sample survey of a "real" population. Accordingly, the weighted least squares (WLS) regression methods were used to estimate models whose cumulative distribution functions (CDFs) came as close as possible to the nominal probabilities at the tabulated percentiles. The models used were the generalized gamma and its three two-parameter special cases (gamma, lognormal, and Weibull). The adequacy of fit of the two-parameter models was evaluated by comparison with the fit of the generalized gamma distribution, using an F test with one degree of freedom for the numerator and three degrees of freedom for the denominator. This is a GOF test relative to the three-parameter model.

4.2 Results

Table 4-3 summarizes results. For each age group, the best fitting two-parameter model is indicated in column 2. Columns 3 through 8 contain the values of the estimated CDFs for these models at the tabulated quantiles from EFH Table 14-159. As for tap water consumption, the goal is to estimate the CDF in order to come as close as possible to the nominal cumulative probabilities of 0.25, 0.50, 0.75, 0.90, 0.95, and 0.99. Columns 9 and 10 contain the estimated mean for the fitted model and the simulated mean from EFH Table 14-159. The next to last column contains the F test *p*-value, PGOF, for goodness-of-fit (GOF) of the selected model relative to the three-parameter generalized gamma model. The generalized gamma distribution improves significantly on the best fitting two-parameter model at the 5% significance level whenever PGOF <0.05. This occurs in 6 of 30 cases. In 20 of 30 cases, the best fitting two-parameter model was the Weibull model.

4.3 Uncertainty Analysis

Information on parameter uncertainty distributions can be summarized as for tap water consumption in Section 3, using parameter estimates and the asymptotic covariance matrix produced by the SAS nonlinear regression (NLIN) procedure. For the gamma and Weibull models, logarithms of the usual positive parameters should be used. For the lognormal model, the parameters should be the mean and logarithm of variance of the logarithm of residence time.

Work is in progress to develop parameter uncertainty distributions for population mobility.

4.4 Conclusions

Given that all three types of the basic two-parameter models are needed to adequately fit the population mobility data, it might appear simpler just to tabulate the best fitting generalized gamma distributions. However, this would somewhat complicate the uncertainty analysis, which would require the use of a trivariate normal distribution with some parameters very highly correlated, or the use of one of the other uncertainty methods.

Another promising approach to population mobility as well as tap water consumption involves the use of generalized gamma regression models (Section 2.2.6).

The analysis of population mobility data focused on the age-specific distributions of EFH Table 14-159 taken from the simulation study of Johnson and Capel (1992). However, Israeli and Nelson (1992) provide results for geographic regions, farms, urban versus rural, and renters versus owners. These factors are also relevant. Efforts are under way within EPA to develop region-specific distributions for residence time.

Extensive information on population mobility is available on the worldwide web (http://www.census.gov/prod/1/pop/p20-485.pdf). We recommend that this information be reviewed to determine its applicability to estimation of population mobility distributions.

Johnson and Capel (1992) developed a methodology to determine the distribution of residential occupancy periods in a simulated population of 500,000 individuals. The input data used for this analysis are considered representative of the general U.S. population. Therefore, the distributions presented in Table 4-3 may be used when the general population is the population of concern and for the age groups presented.

Table 4-1. Selected Percentiles of Residence Times in Years from Three Key Studies

Statistic	Israeli & Nelson (1992)	Census Bureau (1993)	Johnson & Capel (1992)
25th percentile	0.5	4	4
50th percentile	1.4	9	9
75th percentile	3.7	18	16
90th percentile	12.9	32	26
95th percentile	23.1	40	33
Average	4.6	N/A	11.7

Table 4-2. Residence Time[a] Distributions in Years from Johnson and Capel (1992)

Age Group	Mean Years	Percentile					
		25th	50th	75th	90th	95th	99th
00-03	6.50	3	5	8	13	17	22
04-06	8.00	4	7	10	15	18	22
07-09	8.90	5	8	12	16	18	22
10-12	9.30	5	9	13	16	18	23
13-15	9.10	5	8	12	16	18	23
16-18	8.20	4	7	11	16	19	23
19-21	6.00	2	4	8	13	17	23
22-24	5.20	2	4	6	11	15	25
25-27	6.00	3	5	8	12	16	27
28-30	7.30	3	6	9	14	19	32
31-33	8.70	4	7	11	17	23	39
34-36	10.4	5	8	13	21	28	47
37-39	12.0	5	9	15	24	31	48
40-42	13.5	6	11	18	27	35	49
43-45	15.3	7	13	20	31	38	52
46-48	16.6	8	14	22	32	39	52
49-51	17.4	9	15	24	33	39	50
52-54	18.3	9	16	25	34	40	50
55-57	19.1	10	17	26	35	41	51
58-60	19.7	11	18	27	35	40	51
61-63	20.2	11	19	27	36	41	51
64-66	20.7	12	20	28	36	41	50
67-69	21.2	12	20	29	37	42	50
70-72	21.6	13	20	29	37	43	53
73-75	21.5	13	20	29	38	43	53
76-78	21.4	12	19	29	38	44	53
79-81	21.2	11	20	29	39	45	55
82-84	20.3	11	19	28	37	44	56
85-87	20.6	10	18	29	39	46	57
88-90	18.9	8	15	27	40	47	56
Adult	16.2	8	14	22	30	36	48
All	11.7	4	9	16	26	33	47

[a]Number of years between the date that a person moves into a new residence and the date that a person dies or moves out of the residence.

Table 4-3. Results of Statistical Modeling of Population Mobility Data[a]

Age Group	Best Model	P25	P50	P75	P90	P95	P99	Model Mean	Data Mean	PGOF
00-03	Wei2	.287	.485	.709	.904	.965	.991	6.3	6.5	.088
04-06	Wei2	.257	.513	.718	.909	.959	.988	7.7	8.0	.597
07-09	Wei2	.248	.491	.758	.910	.949	.986	8.8	8.9	.299
10-12	Wei2	.217	.527	.779	.894	.940	.989	9.3	9.3	.096
13-15	Wei2	.252	.492	.755	.906	.946	.989	8.8	9.1	.720
16-18	Wei2	.252	.495	.744	.911	.957	.985	8.1	8.2	.745
19-21	Gam2	.260	.480	.752	.905	.956	.987	5.8	6.0	.949
22-24	Log2	.237	.542	.721	.904	.953	.989	5.2	5.2	.782
25-27	Log2	.245	.509	.752	.894	.950	.991	6.4	6.0	1.00
28-30	Log2	.222	.552	.744	.890	.948	.989	7.3	7.3	.336
31-33	Log2	.241	.520	.745	.894	.951	.991	8.9	8.7	.751
34-36	Log2	.259	.495	.739	.901	.953	.991	10.6	10.4	1.00
37-39	Log2	.235	.515	.757	.904	.949	.986	11.8	12.0	.011
40-42	Gam2	.255	.501	.740	.896	.956	.991	13.5	13.5	.192
43-45	Gam2	.248	.513	.732	.905	.953	.989	15.3	15.3	.872
46-48	Wei2	.261	.497	.736	.900	.954	.991	16.4	16.6	.055
49-51	Wei2	.257	.486	.754	.902	.951	.989	17.4	17.4	.882
52-54	Wei2	.241	.501	.759	.903	.951	.987	18.0	18.3	.030
55-57	Wei2	.246	.498	.756	.902	.952	.987	18.9	19.1	.123
58-60	Wei2	.245	.498	.762	.900	.947	.989	19.7	19.7	.204
61-63	Wei2	.235	.517	.749	.904	.949	.988	20.1	20.2	.118
64-66	Wei2	.233	.516	.755	.900	.949	.988	20.9	20.7	.042
67-69	Wei2	.231	.507	.767	.904	.950	.985	21.2	21.2	.002
70-72	Wei2	.253	.493	.755	.896	.952	.990	21.2	21.6	.763
73-75	Wei2	.254	.491	.751	.905	.950	.989	21.7	21.5	.933
76-78	Wei2	.251	.486	.762	.904	.953	.986	21.1	21.4	.413
79-81	Wei2	.229	.520	.754	.904	.951	.986	21.3	21.2	.048
82-84	Wei2	.243	.510	.751	.895	.952	.990	20.6	20.3	.518
85-87	Wei2	.239	.498	.768	.903	.951	.986	20.5	20.6	.039
88-90	Wei2	.244	.483	.770	.919	.956	.981	18.6	18.9	.259

[a]Tabulated probabilities are obtained by evaluating the best fitting two-parameter CDF at the percentile value of Johnson and Capel (1992), from Table 4-2.

5

Application to Inhalation Rates

The treatment of long-term inhalation rates relies on Francis and Feder (1997), who provide a thorough review of available data sources for estimation of long-term and short-term inhalation rate distributions. The authors identify several areas where data are lacking or are out of date and make several recommendations for improving data sources. As Francis and Feder (1997) point out, a potentially important technique for estimating long-term breathing rate distributions is to combine activity pattern distributions with short-term activity-specific breathing rate distributions. Because activity pattern distributions have not yet been developed, this approach is not yet applicable.

Computation of risk for a defined population typically involves sums of products of random variables. The summation is over exposure pathways. For a given pathway, the risk is typically a product of factors. The estimation of distributions for inhalation rates is somewhat similar to a risk assessment, because inhalation rates are themselves the product of component factors.

The approach used to address inhalation rates is considered reasonable when very limited data, such as only estimated means and standard deviations, are available. The method used is motivated by Rai et al. (1996). Assuming independence of the factors within each subpopulation, the mean and standard deviation of the product can be estimated. Since distributional information for the individual factors is not available, model uncertainty is present. Therefore, it is recommended that the lognormal and at least one of the gamma and Weibull distributions with the same estimated product mean and standard deviation be used in risk assessment. In some cases (e.g., if the coefficient of variation [CV] of the product is on the order of 50% or less), these distributions may be sufficiently similar that risk assessment can be reasonably based on any one of them. "For small variance it is likely to be difficult to discriminate between lognormal models and gamma models" (McCullagh and Nelder, 1983).

5.1 Data

Layton (1993) is the only study referenced in the Exposure Factors Handbook (EFH) that allows direct calculation of long-term breathing rate (expressed as cubic meters [m^3] per day) without combining activity distributions with short-term breathing rate distributions. Layton's methods are based on oxygen consumption associated with energy expenditures. The general equation for a metabolically based determination of ventilation rate is:

$$V_E = E * H * VQ \quad (5.1)$$

where

V_E = ventilation rate (inhalation rate) (m^3/day)
E = energy expenditure rate in megajoules/day (MJ/day)
H = oxygen uptake factor, the volume of oxygen (at standard temperature and pressure, dry air) consumed in the production of 1 MJ energy expended (m^3/MJ)
VQ = ventilatory equivalent, the ratio of minute volume to oxygen uptake (unitless).

Layton (1993) presented three approaches for the calculation of V_E based on different methods for estimating the energy expenditure rate. Layton's first method estimates food energy intake from the 1977-1978 National Food Consumption Survey (NFCS) and the National Health and Nutrition Examination Survey (NHANES II). Layton argues that the NFCS and NHANES II estimates are biased low and develops a correction factor of 1.2.

Layton's second method is based on the relationship E=BMR * A, where BMR is the basal metabolic rate (MJ/day) and A is the ratio of the total daily energy expenditure to the daily BMR. BMR values for specific gender/age cohorts are provided by Layton as means and standard deviations that can be used to develop energy expenditure distributions. Sample sizes for each gender/age group range from 38 to 2,879 and are based on Schofield (1985).

Layton's third method is to combine activity pattern distributions with short-term activity-specific breathing rate distributions. For reasons given in the introduction to this section, this approach is

not pursued at this time. The second approach is used here. The first approach is not used because of the arbitrariness of the bias correction factor 1.2.

To apply Layton's second method, the basic equation (5.1) becomes

$$V_E = BMR * A * H * VQ \quad (5.2)$$

where:

BMR = basal metabolic rate (MJ/day)

A = (aka, PAI or MET) ratio of total daily energy expenditure to daily BMR.

Table 5-1 contains the statistical summaries that were used to estimate the distribution of V_E. This is essentially the same information as given by EFH Table 5-12, supplemented by population variance estimates for each quantity. The derivation of the estimated means and standard deviations in Table 5-1 for ventilation rate (V_E) were as follows:

- The mean of oxygen uptake factor (H) was a weighted average from NFCS and NHANES II. We assumed a 10% CV for H based on the idea that any human biochemical attribute must have at least this much variability. However, larger values (10%-20%) may be more reasonable in some situations.

- Ventilatory equivalent (VQ) estimates for ages 0-3 were pooled estimates from Stahlman and Meece (1957) and Cook et al. (1955). A log transformation of the geometric mean and geometric standard deviations reported in the EFH was used to obtain means and standard deviations of log (VQ). A weighted average of the means and variances of log (VQ) was used to obtain pooled estimates. These pooled estimates were then transformed to obtain the mean and variance of an assumed underlying lognormal distribution. (VQ statistics for ages greater than 3 are from Layton, 1993; five studies pooled).

- BMR estimates of the mean were obtained from Table 5-12 in the EFH. EFH Table 5A-4 provided CVs for the same age categories, so the means from Table 5-12 and CVs from Table 5A-4 were used to calculate the standard deviations. Values of BMR and VQ were

chosen to reflect "average" people and are not intended to represent population extremes (e.g., marathon runners).

- Ratio of total daily energy expenditure (A) for ages 0-10 was obtained from Griffiths and Payne (1976). A weighted average of the CVs for A from the 10-60 age groups for males and females to obtain a CV for ages 0-10 was used. Estimates of A for ages 10-60 were obtained from Basiotis et al. (1989). Estimates for ages >60 were obtained from James et al. (1989), who summarized five studies for ages >60. Means were calculated from the three estimates for females and four estimates for males. CVs were assumed to be the same as for ages 10-60.

5.2 Statistical Methods

Since the available data consist of means and standard deviations, the only applicable non-Bayesian estimation technique is the method of moments.

The following calculations were carried out for each of the 12 groups defined by gender and the six age ranges (0-3, 3-10, 10-18, 18-30, 30-60, >60 years). Using the estimated means and standard deviations from Table 5-1, the mean and variance of inhalation rate (BMR*A*H*VQ) were estimated, assuming the four factors are statistically independent within each subpopulation. Since independence is assumed only within each subpopulation, this allows for some dependence among the factors (i.e., does not assume overall independence of these factors). Independence implies that the mean of the product is the product of the means. If X and Y are two independent random variables with means MX, MY, and variances VX, VY, then the variance of the product of X*Y is given by

$$\text{variance of } X*Y = VX*VY + VX*MY^2 + MX^2 * VY.$$

The variance of a product of more than two terms can be obtained by repeatedly applying this relation. Finally, the gamma and lognormal distributions with the given product means and variances were calculated and compared.

Since only moment information was available, the standard goodness-of-fit tests are not applicable.

To obtain parameter distributions for uncertainty analysis, the following two approaches are possible. Under the assumption that each of the four factors has a specific distribution (e.g., lognormal), distributions for the individual factor means and variances could be obtained by using a normal or t distribution for the mean and a chi-square distribution for the variance. However, this would be a questionable approach if gamma distributions were assumed for the factors. A bootstrap method would be applicable for either the gamma or lognormal assumption.

The bootstrap method using the gamma case is described for illustration. One thousand simulated copies of Table 5-1 would be generated by assuming gamma distributions with the tabulated means and standard deviations for each factor and group. For each of these 1,000 tables, the 12 estimated means and variances of the products of Equation 5.2 would be calculated. This would yield a bootstrapped distribution for the gamma parameters.

This uncertainty analysis requires special methods based on the method of moments and is planned for a future manuscript.

5.3 Results

Table 5-2 contains estimated means and CVs for inhalation rates, using the methods of Section 5.2 for calculating the mean and variance of a product of independent random variables. Except for ages <3, most of the CVs are on the order of 30%. Because the CVs are of moderate size, the quantiles of the estimated gamma (XG50, XG90, XG95, XG99) and lognormal (XL50, XL90, XL95, XL99) distributions are reasonably similar. The %Diff measure was calculated as the average absolute percent discrepancy between the four gamma and lognormal estimated quantiles, that is, as $|XG99 - XL99| / [(XG99 + XL99) / 2]$, using the average of the two estimated quantiles as the nominal value.

5.4 Conclusions

For most purposes, the difference between the gamma and lognormal distributions would probably be negligible, and the lognormal distributions could be used. If the CVs had been larger (e.g., CV>100%), the difference between gamma and lognormal distributions would have been much more

evident. It thus appears that lack of knowledge of distributional form for the individual factors is not a serious drawback, because the individual factors and the product do not have large CVs.

However, the values assumed for the means and standard deviations of individual factors are important determinants of the product distribution. Francis and Feder (1997) have reviewed the available data sources, including those used here and have made a number of recommendations for updating and improving estimates based on more recent and more relevant data.

The results presented in this analysis are based on the Layton (1993) study in which inhalation rates were determined indirectly, based primarily on the BMR and energy expenditures. BMR values were determined based on the literature, and energy expenditures were calculated based on the USDA 1977-78 NFCS. Therefore, this distribution may be used when conducting an assessment where the U.S. national population is of concern. These values also represent daily average inhalation rates and are not applicable to activity-specific inhalation rates (short-term).

Table 5-1. Parameter Estimates for Individual Factors Affecting Long-Term Inhalation Rates (m^3/day)

Parameter	Age Group	Gender	Mean	Standard Deviation	Coefficient of Variant	Sample Size
H	ALL	Both	0.05	0.005	10.0	51,092
VQ	0-3	Both	28.01	7.44	26.6	61
VQ	>3	Both	27.37	4.56	16.7	75
BMR	0-3	Male	3.40	2.07	60.9	162
BMR	3-10	Male	4.30	0.52	12.1	338
BMR	10-18	Male	6.70	1.34	20.0	734
BMR	18-30	Male	7.70	0.92	11.9	2,879
BMR	30-60	Male	7.50	0.98	13.1	646
BMR	>60	Male	6.10	1.04	17.0	50
BMR	0-3	Female	2.60	1.53	58.8	137
BMR	3-10	Female	4.00	0.52	13.0	413
BMR	10-18	Female	5.70	0.86	15.1	575
BMR	18-30	Female	5.90	0.83	14.1	829
BMR	30-60	Female	5.80	0.64	11.0	372
BMR	>60	Female	5.30	0.64	12.1	38
A	0-10	Both	1.58	0.30	19.6	12
A	10-60	Male	1.59	0.33	20.8	13
A	10-60	Female	1.38	0.24	17.4	16
A	>60	Male	1.52	0.32	20.8	14
A	>60	Female	1.44	0.25	17.4	14

Table 5-2. Estimated Mean, Coefficient of Variation, and Quantiles for Inhalation Rate (m^3/day), Assuming Gamma or Lognormal Distribution

Age	Sex	Mean	CV	XG50	XL50	XG90	XL90	XG95	XL95	XG99	XL99	%Diff
00-03	M	7.52	73	6.2	6.1	14.9	14.1	18.2	17.8	25.7	27.9	4.61
00-03	F	5.75	71	4.8	4.7	11.2	10.7	13.7	13.4	19.2	20.8	4.47
03-10	M	9.30	30	9.0	8.9	13.0	13.0	14.3	14.5	17.0	17.7	1.54
03-10	F	8.65	31	8.4	8.3	12.2	12.1	13.4	13.5	16.0	16.6	1.57
10-18	M	14.58	36	14.0	13.7	21.5	21.4	24.0	24.2	29.2	30.6	1.94
10-18	F	10.76	31	10.4	10.3	15.1	15.1	16.7	16.8	19.9	20.7	1.57
18-30	M	16.75	31	16.2	16.0	23.7	23.7	26.2	26.5	31.3	32.6	1.63
18-30	F	11.14	30	10.8	10.7	15.6	15.6	17.2	17.3	20.4	21.2	1.53
30-60	M	16.32	32	15.8	15.6	23.2	23.2	25.7	25.9	30.8	32.0	1.66
30-60	F	10.95	29	10.7	10.5	15.1	15.1	16.6	16.7	19.6	20.3	1.43
>60	M	12.69	34	12.2	12.0	18.4	18.4	20.5	20.7	24.8	25.9	1.83
>60	F	10.44	29	10.2	10.0	14.5	14.5	15.9	16.0	18.8	19.5	1.46

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. Any other approach entails questionable assumptions of independence of factors.

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 data are so limited that most scientists and statisticians would prefer not to make distributional assumptions. 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.

7

References

Barndorff-Nielsen, OE; Cox, DR. (1994) Inference and asymptotics. New York: Chapman and Hall.

Basiotis, PP; Thomas, RG; Kelsay, JL; Mertz, W. (1989) Sources of variation in energy intake by men and women as determined from one year's daily dietary records. Am J Clin Nutr 50:448-452.

Bowman, AW; Azzalini, A. (1997) Applied smoothing techniques for data analysis. Oxford, U.K: Clarendon Press, 193 pp.

Canadian Ministry of National Health and Welfare. (1981) Tapwater consumption in Canada. Document number 82-EHD-80, Public Affairs Directorate, Department of National Health and Welfare, Ottawa, Canada.

Chambers, JM. (1973) Fitting nonlinear models: numerical techniques. Biometrika 60(1):1-13.

Conover, WJ. (1980) Practical nonparametric statistics. New York: John Wiley and Sons, Inc., 493 pp.

Cook, CD; Cherry, RB; O'Brien, D; Kalber, P; Smith, CA. (1955) Studies of respiratory physiology in the newborn infant. 1. Observations on normal and premature and full-term infants. J Clin Invest 34:975-982.

Cox, DR. (1990) Role of models in statistical analysis. Stat Sci 5(2):169-174.

D'Agostino, R; Stephens, MA,eds. (1986) Goodness-of-fit techniques. New York: Marcel Dekker, Inc.

Dixon, PM. (1993) The bootstrap and the jackknife: describing precision in ecological studies. In: Design and analysis of ecological experiments. Scheiner, SM; Gurevitch, J, eds. New York: Chapman and Hall, 445 pp.

Efron, B. (1982) Maximum likelihood and decision theory. Ann Stat 10(2):340-356.

Efron, B; Gong, G. (1983) A leisurely look at the bootstrap, the jackknife and cross-validation. Am Stat 37(1):36-48.

Efron, B; Tibshirani, R. (1993) An introduction to the bootstrap. New York: Chapman and Hall.

Ershow, AG; Cantor, K. (1989) Total water and tapwater intake in the United States: population-based estimates of quantities and sources. Life Sciences Research Office Monograph. Bethesda, MD: Federation of American Societies for Experimental Biology. Available from 9650 Rockville Pike, Bethesda, MD 20814.

Evans, M; Hastings, N; Peacock, B. (1993) Statistical distributions, 2nd ed. New York: J. Wiley and Sons, Inc., 170 pp.

Francis, M; Feder, P. (1997) Development of long-term and short-term inhalation rate distributions. Draft report. Battelle Memorial Institute, Columbus, Ohio.

Griffiths, M; Payne, PR. (1976) Energy expenditure in small children of obese and nonobese parents. Nature 260:698-700.

Hattis, D; Burmaster, D. (1994) Assessment of variability and uncertainty distributions for practical risk analyses. Risk Anal 14(5):713-730.

Israeli, M; Nelson, CB. (1992) Distribution and expected time of residence for U.S. households. Risk Anal 12(1):65-72.

James, WPT; Ralph, A; Ferro-Luzzi, A. (1989) Energy needs of elderly, a new approach. In: Munro, HN; Danford, DE, eds. Nutrition, aging and the elderly. New York: Plenum Press, pp. 129-151.

Jennrich, RI; Ralston, ML. (1979) Fitting nonlinear models to data. Ann Rev Biophys Bioeng 8:195-238.

Johnson, NL. (1978) Approximations to distributions. In: International encyclopedia of statistics. Kruskal, WH; Tanur, JM, eds. New York: The Free Press, a division of the Macmillan Company.

Johnson, NL; Kotz, S. (1970) Continuous univariate distributions, vols. 1 and 2. New York: John Wiley and Sons, Inc.

Johnson, T; Capel, J. (1992) A Monte Carlo approach to simulating residential occupancy periods and its application to the general U.S. population. Research Triangle Park, NC: U.S. Environmental Protection Agency, Office of Air Quality and Standards.

Kalbfleisch, JD; Prentice, RL. (1980) The statistical analysis of failure time data. New York: John Wiley and Sons, Inc.

Kendall, MG; Buckland, WR. (1971) A dictionary of statistical terms, 3rd ed. New York: Hafner Publishing Co., Inc., 166 pp.

Kendall, M; Stuart, A. (1979) The advanced theory of statistics (three volumes). New York: Macmillan Publishing Company, Inc.

Kleinbaum, DG; Kupper, LL; Muller, KE. (1988) Applied regression analysis and other multivariable methods. Boston: PWS-Kent, 718 pp.

Kotz, S; Johnson, NL. (1985) Encyclopedia of statistical sciences (9 volumes and index). New York: John Wiley and Sons, Inc.

Krieger, AM; Pfeffermann, D. (1997) Testing of distributions from complex surveys. J Off Stat 13(2):123-142.

Kruskal, W; Mosteller, F. (1979) Representative sampling I: non-scientific literature. Intern Stat Rev 47:13-24.

Kruskal, W; Mosteller, F. (1979) Representative sampling II: scientific literature. Intern Stat Rev 47:111-127.

Kruskal, W; Mosteller, F. (1979) Representative sampling III: the current statistical literature. Intern Stat Rev 47:245-265.

Law, AM; Kelton, WD. (1991) Simulation modeling and analysis, 2nd ed. New York: McGraw-Hill, Inc.

Layton, DW. (1993) Metabolically consistent breathing rates for use in dose assessments. Health Phys 64(1):23-36.

Lehman, HJ. (1994) Homeowners relocating at faster pace. Virginia Homes Newspaper, Saturday, June 15, p. E1.

McCullagh, P; Nelder, JA. (1983) Generalized linear models. New York: Chapman and Hall.

Mendenhall, W; Wackerly, DD; Scheaffer, RS. (1990) Mathematical statistics with applications, 4th ed. Boston: PWS-Kent, 713 pp.

Millard, SP. (1998) EnvironmentalStats for S-Plus. New York: Springer, 381 pp.

Millard, SP; Neerchal, NK. (1999) Environmental statistics. *In press.* Boca Raton, FL: Chapman and Hall/CRC, 416 pp.

National Association of Realtors. (1993) The estate business series. Washington, DC: National Association of Realtors, Palisade Corporation.

Prentice, R. (1975) Discrimination among some parametric models. Biometrika 62(3):607-614.

Press, WH; Teukolsky, SA; Vetterling, WT; Flannery, BP. (1992) Numerical recipes in C: the art of scientific computing. Cambridge, UK: Cambridge University Press.

Rai, SN; Krewski, D; Bartlett, S. (1996) A general framework for the analysis of uncertainty and variability in risk assessment. Hum Ecol Risk Assess 2(4):972-989.

Schofield, W. (1985) Predicting basal metabolic rate, new standards, and reviews of previous work. Hum Nutr Clin Nutr 39C(suppl) 1:5-41.

Serfling, R. (1980) Approximation theorems of mathematical statistics. New York: John Wiley and Sons, Inc.

Shah, BV; Barnwell, BG; Bieler, GS. (1997) SUDAAN user's manual, release 7.5. Research Triangle Park, NC: Research Triangle Institute. (email: sudaan@rti.org; website: http://www.rti.org/patents/sudaan/html)

Stahlman, MT; Meece, NJ. (1957) Pulmonary ventilation and diffusion in the human newborn infant. J Clin Invest 36:1081-1091.

Stephens, MA. (1974) EDF statistics for goodness of fit and some comparisons. J Am Stat Assoc 69(347):730-737.

Thompson, SK. (1992) Sampling. New York: John Wiley and Sons, Inc., 343 pp.

U.S. Bureau of the Census. (1993) Geographical mobility: March 1991 to March 1992. Current Population Reports, pp. 20-473.

U.S. Department of Agriculture. (1984) Nationwide Food Consumption Survey, 1977-78 Individual Intake Data. Spring, Summer, Fall and Winter Basic Individual Food Intake Surveys. NTIS Accession Nos. PB80190218/HBF, PB80-197429/HBF, PB8020023/HBF, and PB81-118853/HBF. Springfield, VA: National Technical Information Service, US Department of Commerce.

U.S. Environmental Protection Agency. (1997a) Office of Research and Development, National Center for Environmental Assessment, Washington, DC. Exposure factors handbook. Final, vols I, II, III. EPA/600/P-95/002F(a-c).

U.S. Environmental Protection Agency. (1997b) Guiding principles for Monte Carlo analysis. Risk Assessment Forum, EPA/630/R-97/001.

U.S. Environmental Protection Agency. (1999) Report of the Workshop on Selecting Input Distributions for Probabilistic Assessments. National Center for Environmental Assessment, Washington, DC. EPA/630/R-98/004.

This page intentionally left blank.

Appendix A

Glossary

asymptotic normality—Refers to the condition in which the sampling distribution of a parameter estimate approaches that of a normal distribution as the sample size becomes "large." Depending on the estimator, large usually means 30 to 60 observations. When these conditions hold, the estimate is said to be asymptotically normal, and the normal approximation can be used to establish confidence limits for the parameters. One of the many desirable attributes of maximum likelihood estimators is that they are asymptotically normal under fairly simple but broadly applicable conditions.

Bayesian inference—A method that regards model parameters as random variables with prior probability distributions reflecting prior knowledge about the parameters. Bayesian inference is based, via Bayes Theorem, on the conditional (posterior) distribution of the parameters, given the data.

bootstrap estimation—A technique for estimating the variance and/or the bias of a sample estimate of a population parameter by repeatedly drawing (with replacement) a large number (e.g., 1,000) of new, "bootstrap" samples from the original sample. The sample size of each bootstrap sample is the same as the original sample. The variance and bias estimators are computed from the distribution of the bootstrap samples. This technique is most useful for cases where there is no known closed-form estimator for the population variance or in other situations where the usual estimators are not appropriate (e.g., for small sample sizes).

coefficient of variation (CV)—A dimensionless measure of dispersion, equal to the standard deviation divided by the mean, often expressed as a percentage.

complex sampling design—A sampling design in which individual population elements do not have equal probabilities of selection. Complex sample surveys generally incorporate stratification and/or clustering wherein the population members may be correlated. As a consequence, the iid assumption (see p. A-3) may not hold for the sampled population members.

confidence interval—The interval or region about a sample estimate within which the desired population parameter is expected to occur with some specified probability (i.e., the true value of the population parameter will lie within the interval or range for 95% of all samples).

continuous random variable—A random variable that may take on an infinite number of values. The cumulative distribution function of a continuous random variable is therefore a smooth function.

correlation coefficient—A scale-invariant measure of the association between two variables that takes on values between –1 and +1. The correlation coefficient has a value of +1 whenever an increase in one is accompanied by an increase in the other, zero when there is no relationship (i.e., the two variables are independent of one another), and -1 when there is an exact inverse relationship between them.

covariance—A scale-dependent measure of the tendency of the values of one variable to change with those of a second variable. Algebraically, the covariance is the expected value of the product of the deviations of two random variables from their respective means. When this product is zero, the two variables are said to be uncorrelated; otherwise, they will be correlated.

covariates—Random variables (discrete and/or continuous) that are specified as predictor variables in a multivariable model.

cumulative distribution function (CDF)—F(x) equals the probability that a randomly chosen member of a population has a value less than or equal to x for the variable of interest. With reference to a random variable X, the CDF of X, F(x), is the probability that the random variable X does not exceed the number x. Symbolically, $F(x) = P[X \leq x]$.

degrees of freedom (df)—As used in statistics, df has several interpretations. A sample of n variate values is said to have n degrees of freedom, but if k functions of the sample values are held constant, the number of degrees of freedom is reduced by k. In this case, the number of degrees of freedom is conceptually the number of independent observations in the sample, given that k functions are held constant. By extension, the distribution of a statistic based on n independent observations is said to have n-p degrees of freedom, where p is the number of parameters of the distribution.

discrete random variable—A random variable that may take on only a finite number of values. The CDF of a discrete random variable is therefore a step function.

empirical distribution function (EDF)—The sample estimate of the CDF. For any value of $X=x_i$, it is the proportion of observations that are less than or equal to x_i. The graph of the EDF is a step function for which the value at $X=x_i$ is n_i/n, where n_i is the number of sample observations with values of $X \leq x_i$ and n is the total number of observations in the sample. The plot is a series of steps ascending, left to right, from 0 to 1.

goodness-of-fit (GOF) test—Any of several statistical tests of the null hypothesis that the population distribution of the observations is a specified probability distribution or is in a specified set of probability distributions (e.g., the lognormal distribution). The tests evaluate whether or not the EDF is significantly different from the specified CDF.

iid assumption (independent and identically distributed assumption)—Assumes that the values of a random variable in a sample are not correlated with each other and that they share a common probability density function (PDF). This assumption will hold for data collected by simple random sampling but (usually) not for data from more complex (i.e., stratified and/or clustered) sampling designs.

kernel density estimation—A technique for estimating the probability density of a distribution by fitting a smooth curve to the underlying frequency histogram. The choice of the degree to which the distribution should be smoothed is crucial and usually is based on criteria that minimize the mean square error. Unlike the parametric methods that depend on the parameters of a known theoretical PDF, the kernel estimate is derived entirely from the attributes of the sample EDF.

key study—Designation used in the Exposure Factors Handbook (U.S. EPA, 1997a) to distinguish the studies that were regarded as the most useful (representative) for deriving a recommendation for an exposure factor.

likelihood ratio test (LRT)—A parametric test of a null hypothesis that uses as its test statistic (-2) times the natural logarithm of the ratio of two maximized likelihoods. The numerator likelihood is maximized under the constraint (condition) of the null hypothesis. The denominator likelihood does not

have this constraint. The test statistic is usually assumed to have a chi-squared distribution with degrees of freedom equal to the differences in dimensionality of the two parameter spaces.

maximum likelihood estimator (MLE)—The parameter estimates that maximize the probability of obtaining the sample observations.

meta-analysis—The process of using statistical techniques to combine the results of several different studies. Meta-analyses may permit stronger and/or broader inferences than were possible in any of the constituent studies.

model parameter—Numerical characteristic of a given population (e.g., the mean and variance of a normal population) that determines some response of interest in accordance with a specific mathematical formula. Such an expression is called a model. By convention, statistical model parameters are usually symbolized as Greek letters.

Monte Carlo methods—Methods used to investigate the properties of an inferential procedure by applying it to computer-generated data that serve as a surrogate for "real data" collected by random sampling.

multivariate parametric distribution—The joint theoretical probability distributions of two or more random variables. The component univariate distributions can be of the same kind (e.g., three lognormal distributions), or they may be combinations of several different kinds (e.g., lognormal, Weibull, and exponential). Typically, the component variables are correlated; thus, correlations comprise additional parameters of multivariate distributions.

P-P (probability-probability) plot—A graph used to subjectively assess GOF. For any given $X=x_i$, the value of the CDF for the theoretical distribution of interest is plotted on one axis, and the observed value of the EDF for $X=x_i$ is plotted on the other axis. P-P plots that closely approximate a diagonal line through the origin indicate a good fit between the EDF and the theoretical CDF.

***p*-value**—A value between 0 and 1 that is often regarded as a measure of the belief in a statistical null hypothesis (H_0). In the Frequentist view (vs. the Bayesian view), a test statistic has a specific parametric distribution when H_0 is true. A test statistic is computed from sample data and compared with its

expected parametric distribution. The *p*-value is the probability that a value of the test statistic, as extreme or more extreme than the observed test statistic, came from the null distribution. If the *p*-value is less than or equal to the significance level (α) of the test, the null hypothesis is rejected. A major objection of Bayesian statisticians to the Frequentist approach is that the specification of the α value (usually 0.05) is arbitrary.

percent error plots—A graphical GOF test. The difference between the observed and hypothesized quantile values, expressed as a percent of the hypothesized value, is plotted on the vertical axis versus the observed quantiles on the horizontal axis. Because the points on the plot are compared with a horizontal reference line (i.e., percent difference=0) and because of the relative nature of the differences being displayed, lack-of-fit is more apparent than in P-P or Q-Q plots.

point mass at zero—A positive probability that the observed value of the random variable is zero (e.g., the probability that the amount of tap water consumed by an infant per day is zero). If the distribution of a random variable, X, is a continuous parametric distribution, the probability of observing a value in the interval from X=a to X=b is equal to the area under the probability density function (the derivative of the CDF) between a and b. By definition, a single point (e.g., X=0) does not occupy any space and, hence, has probability zero of occurring exactly. Therefore, the distribution for some exposure factors may be a composite probability distribution that includes a positive probability of observing X=0 and a continuous parametric distribution (e.g., lognormal) for positive values of X.

probability density function (PDF)—The PDF of a continuous random variable X is the first derivative of its CDF, $f(x) = F'(x)$. The probability that $a \leq X \leq b$ is found by integrating f(x) from a to b.

Q-Q plot—A graph used to subjectively assess GOF. For any given probability value p_i, the value of the random variable, X, for which the theoretical CDF is p_i is plotted on one axis, and the observed value of $X=x_i$, for which the EDF is p_i is plotted on the other axis. Q-Q plots that closely approximate a diagonal line through the origin indicate a good fit between the EDF and the theoretical CDF. Depending on observed patterns of deviation from the diagonal, lack-of-fit due specifically to differences in location (i.e., mean or median) and/or scale (i.e., variance) can be diagnosed.

quantile—The q-1 partition values of a random variable that divide a sample or population into q subdivisions, each of which contain an equal proportion of the sample or population. For example, when

q=4, the three resulting values are the first, second (=median), and third quartiles that collectively divide the data into four equal parts.

random variable—A numeric event whose values change from one sampling unit or one experimental unit to the next. Random variable values may be either discrete or continuous.

relevant study—Designation in the Exposure Factors Handbook (U.S. EPA, 1997a) to distinguish the studies that were applicable or pertinent, but not necessarily the most important for making a recommendation for an exposure factor.

representative sample—A sample that captures the essence of the population from which it was drawn; one which is typical with respect to the characteristics of interest, regardless of the manner in which it was chosen. While representativeness in this sense cannot be completely assured, randomly selected samples are more likely to be representative than are haphazard or convenience samples. This is true because only in random sampling will every population element have an equal probability of selection.

residence time—The time in years between a person moving into a residence and the time the person moves out or dies.

risk assessment—Qualitative or quantitative estimation of the probability of adverse health or environmental effects due to exposure to specific behavioral, dietary, environmental, occupational, or social factors.

sensitivity analysis—The process of varying one or more model parameters while leaving the others constant to determine their effect on the model predictions. The results help to identify the variables that have the greatest effect on model estimates and may be useful for fine-tuning the model or identifying problems for additional research.

simple random sampling design—A sampling design in which every member of the target population has an equal probability (p=1/N) of selection to the sample. Random variables measured on observations from a simple random sample satisfy the iid assumption.

tap water—Water consumed directly from the tap as a beverage or used in preparation of foods and beverages (coffee, tea, frozen juices, soups, etc.).

uncertainty analysis—Identification of the components of variability of risk that are due to model uncertainty or parameter uncertainty, that is, to uncertainty in the type of model (e.g., gamma vs. lognormal vs. Weibull) or uncertainty in the values of model parameters. Parameter uncertainty can be built into risk assessment simulations by randomly drawing population parameters from appropriate distributions before selecting individuals from the population. Model uncertainty can be addressed by sensitivity analysis, using separate simulations for different viable competing models.

univariate parametric distribution—A theoretical probability distribution for a random variable whose CDF is described by a mathematical function of population parameters (e.g., the population mean and variance), such as a normal or lognormal distribution.

Appendix B

Fitting Models to Percentile Data

The Exposure Factors Handbook (EFH) (U.S. EPA, 1997a) often uses percentiles to summarize data for an exposure factor. Let x denote the random variable of interest, that is, x=daily tap water consumption or x=daily inhalation rate. Theoretically, the 100*p*th percentile of a continuous distribution with cumulative distribution function (CDF) F(x) is the value x_p for which $F(x_p)=p$. That is, the 100pth percentile is the value x_p for the variable of interest that places $100p\%$ of the probability below x_p.

A precise definition for empirical percentiles is rather involved because of finite sample size complications. If the sample size is large enough, think of the 100pth percentile simply as the smallest data value (x_p) with at least $100p\%$ of the sample below it. It can be estimated from the linearly interpolated empirical distribution function (EDF) by reading over from p on the vertical axis to the graph of the linearized EDF, then dropping straight down to the horizontal axis to obtain x_p.

The EDF contains all the information in the sample. Ideally, raw data would be available, and we could calculate and work with the EDF. However, raw data often is unavailable because the published literature rarely provides it. Even if raw data are available, it is not practical to include all data points for large samples in the EFH. A summary of percentiles such as those corresponding to p=0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95, and 0.99 contains much of the information in the original data and can be used as a basis for estimation of the distribution and testing goodness-of-fit (GOF).

A variety of methods for fitting distributions to percentile data can be identified. Four are discussed, and three of them are illustrated with a drinking water example from the EFH.

The problem of estimating distributions for exposure factors seems complicated enough by the fact that more than a dozen families of theoretical probability distributions may be needed in a toolkit for fitting environmental data. The most credible and widely used fitting method is maximum likelihood (ML) estimation. Why not simply use ML estimation? Because it may not be the best method. Some evidence of this is shown in the treatment of the tap water consumption data in Section 3.

B.1 Four Methods of Fitting Parametric Models to Percentile Data

Serfling (1980) provides procedures for statistical inference for quantiles based on a large sample.

We concentrate here on three methods that have better small sample properties, which basically select an estimated distribution by attempting to make the fitted probabilities $F(x_p)$ close to the nominal values of 0.01, 0.05, 0.10, etc. Graphically, the data are summarized as a plot of the nine points with x_p plotted on the horizontal axis and p plotted on the vertical axis. The goal is to find a theoretical model that passes close to the nine data points. The three methods are obtained by using different notions of closeness and are referred to as weighted least squares (WLS), minimum chi-square (MCS), and ML approaches.

EXAMPLE: Calculation of WLS, MCS, and ML measures for the tap water consumption data of older adults.

This example is from Table 3-7 of the EFH. The empirical quantile values x_p have the property that $100p\%$ of the sample are below them. The values of x_p and p are in columns 3 and 4 of Table B-1. The quantile values x_p in Table B-1 are those from Table 3-7 divided by 100. This rescaling improves the performance of iterative search methods used to fit the curves.

The results in Table B-1 are from fitting a gamma distribution. The notes for Table B-1 indicate how the various columns are calculated. Column 5 contains the estimated or fitted probabilities $F(x_p)$. The goal of fitting is to choose F to make these $F(x_p)$ values close to the target p's. This gamma distribution was chosen to minimize a weighted sum of squares of errors (WSE) whose individual terms are

$$n*[F(x_p)-p]*[F(x_p)-p]/[p*(1-p)].$$

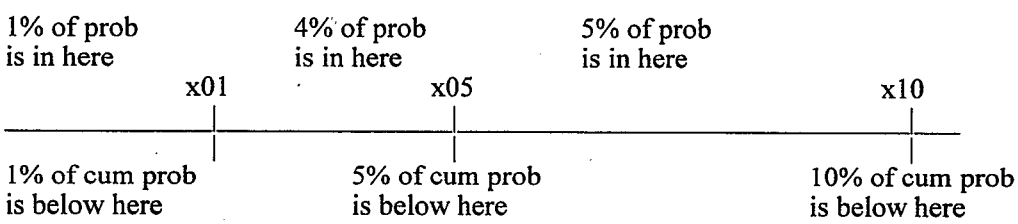
These terms are given in column 6 of Table B-1, labeled "Wtd Sqd Err (WSE)." For example, the WSE term corresponding to p=0.50 is

$$2541*[(.5 - .4942)*(.5 - .4942)]/[(.5)*(.5)] = .345.$$

The column total 13.57 is the minimized WSE. That is, F was chosen as the gamma distribution, which minimizes the sum of these nine WSE terms.

By comparison with the defining formula for the Anderson-Darling (AD) statistic (Law and Kelton, 1991), it can be seen that this WSE measure is the AD discrepancy limited to the nine available quantiles. Intuitively, if a parametric distribution that agrees closely with the data at the available quantiles is selected, good agreement with respect to any aspect of the distribution, such as the mean, should be obtained.

The chi-square and log-likelihood values for this particular fitted model also are calculated on the right-hand side of Table B-1. Unlike the WSE/AD measure, the chi-square and likelihood measures focus on individual rather than cumulative probabilities associated with intervals. This distinction is illustrated in the diagram below.



Thus, column 7 of Table B-1 for nominal probability mass (labeled "Nom Prob Mass pm") contains successive differences between the nominal cumulative probability values. Similarly, column 8 for estimated probability mass (labeled "Estd Prob Mass pm^") contains successive differences between the gamma estimated cumulative probability values $F(x_p)$. The observed and expected numbers (O and E) of sample points in each interval are the products of the sample sizes times these nominal and estimated individual probabilities. That is, column 9 is the product of column 2 times column 7, and column 10 is the product of column 2 times column 8. The chi-square values in column 11 are calculated as (O-E)*(O-E)/O. The first chi-square value is (25.41-9.57)*(25.41-9.57)/25.41 = 9.874. The log-likelihood values are the natural logarithms of pm^ raised to the O power, that is, O*log(pm^).

The sum of the chi-square and log-likelihood values for the fitted gamma distribution are 17.60 and -4870. To obtain the MCS and ML solutions, the gamma parameters would be selected to minimize the chi-square or maximize the likelihood, rather than to minimize the WSE measure.

Table B-1. Example Calculations of Criteria Functions Using Tap Water Consumption Data for Adults Over Age 65 from Table 3-7 of EFH

1 Age Group	2 Sample Size n	3 Quantile x_p	4 Nom Cum prob p	5 Gamma Estd F (x_p)	6 Wtd Sqd Err (WSE)	7 Nom Prob Mass pm	8 Estd Prob Mass pm^	9 Obsd Num O	10 Expd Num E	11 Chi-Square	12 Log Like
65+	2541	0.045	0.01	0.0038	9.974	0.01	0.0038	25.41	9.570	9.874	-142
65+	2541	0.087	0.05	0.0497	0.005	0.04	0.0459	101.6	116.7	2.232	-313
65+	2541	0.109	0.10	0.1051	0.730	0.05	0.0554	127.1	140.7	1.477	-368
65+	2541	0.150	0.25	0.2588	1.041	0.15	0.1537	381.2	390.5	0.229	-714
65+	2541	0.203	0.50	0.4942	0.345	0.25	0.2354	635.3	598.2	2.165	-919
65+	2541	0.271	0.75	0.7425	0.767	0.25	0.2483	635.3	631.0	0.029	-885
65+	2541	0.347	0.90	0.8989	0.031	0.15	0.1565	381.2	397.6	0.708	-707
65+	2541	0.400	0.95	0.9516	0.134	0.05	0.0526	127.0	133.8	0.354	-374
65+	2541	0.513	0.99	0.9915	0.540	0.04	0.0399	101.6	101.3	0.001	-328
65+	2541		1.0	1.000		0.01	0.0085	25.41	21.73	0.534	-121
					13.57					17.60	-4870

Notes: Sample calculations for row 3, p=0.10.
Column 3 = the empirical quantile from EFH Table 3-7, divided by 100.
Column 5 = fitted gamma CDF value $F(X_p)$.
Column 6 = $n^*\{[p\text{-}F(x_p)]^*[p\text{-}F(x_p)]\}/[p(1\text{-}p)]$, using columns 2, 4, 5.
Column 7 = nominal probability mass=p - (last p)=0.10-0.05=0.05.
Column 8 = estimated prob mass = $F(x_p)$- [last $F(x_p)$] = 0.1051-0.0497.
Column 9 = observed number O = n*pm = 2541*(0.05) = 127.1.
Column 10 = expected number E = n*pm^ = 2541*(0.0554) = 140.7.
Column 11 = chi-square = [(O-E)*(O-E)]/O.
Column 12 = log-likelihood = O*log(pm^) = 127.1*(-2.893) = -368.

Appendix C

Fitting Quantiles by Combining Nonlinear and Linear Regression

The approach outlined below was motivated by a July 3, 1997, memorandum from Timothy Barry, Senior Analyst, Office of Policy and Re-Invention, to Jackie Moya, Environmental Engineer, Office of Research and Development.

Let F(x) be the cumulative distribution function (CDF) of a nonnegative continuous random variable X, that is, F(x)=P[X≤x]=the probability of a value≤x. Since X is continuous, F is continuous and strictly increasing, and its inverse FINV exists, so that F[FINV(p)]=p and FINV[F(x)]=x. Let Y=a X^r be a power transform of X with both a and r strictly positive (a>0, r>0), and let G(y)=P[Y≤y] be the CDF of Y. Recall that y_p is the p th quantile of Y iff $G(y_p)=p$. Here iff denotes logical equivalence ("if and only if").

Using basic algebra, set theory, and probability, it can be shown that

$$\log(y_p)=r \cdot \log[\text{FINV}(p)]+\log(a). \qquad \text{(C.1)}$$

Hence, if F and its inverse FINV are known, and there are empirical quantiles y_p for several different values of p, then the power transform parameters a and r by linear regression of log (y_p) on log [FINV(p)] can be estimated. This is easily extended to cover distributions that are nonnegative and continuous except for a point mass M at zero. To see this, let H(y)=0 for y<0, H(y)=M+(1-M) G(y) for y≥0, and note that $H(y_p)=p$ iff $G(y_p)=(p\text{-}M)/(1\text{-}M)$. Hence for p>M, the pth quantile y_p for H is obtained by solving $G(y_p)=p_1$, where $p_1=(p\text{-}M)/(1\text{-}M)$. This leads to

$$\log(y_p)=r \cdot \log[\text{FINV}(p_1)]+\log(a). \qquad \text{(C.2)}$$

These arguments suggest the following combined nonlinear/linear regression approach to fitting the five-parameter generalized F distribution with a point mass M at zero.

Let pmin be the smallest p for which a positive empirical quantile y_p exceeds zero. Then M should not exceed $p$min.

1. Perform an outer search on M, or simply use a grid of M values, such as
   M = 0, 0.1 $p$min, 0.2 $p$min, . . ., 0.9 $p$min.

2. For a given value of M, perform a two-dimensional search on the degrees-of-freedom parameters df_1, df_2 of the generalized F distribution.

3. Given M, df_1, and df_2, estimate a and r by solving the linear regression problem defined by Equation C.2.

Appendix D

The Generalized (Power Transformed) F Family of Nonnegative Probability Distributions

Probability Distribution	**Probability Density Function (PDF)**	**Cumulative Distribution Function (CDF)**	***r*th moment E(Xr)**
General X with parameter vector θ	$f_X(x\|\theta) = F'_X(x\|\theta)$ or $f_X(x) = F'_X(x)$(θ suppressed)	$F_X(x\|\theta) = Prob[X \le x\|\theta]$	$\mu_X(r,\theta) = \int_0^\infty x^r f_X(x\|\theta)\,dx$
Monotonic transformation $Y = g(X)$, $X = h(Y) = g^{-1}(Y)$	$f_Y(y\|\theta) = f_X[h(y)\|\theta]\|h'(y)\|$	$F_Y(y\|\theta) = F_X[h(y)\|\theta]$ for h increasing $F_Y(y\|\theta) = 1 - F_X[h(y)\|\theta]$ for h decreasing	$\int_{g(0)}^{g(\infty)} y^r f_Y(y)\,dy$ $= \int_0^\infty g(x)^r f_X(x)\,dx$
$Y = 1/X$	$f_Y(y) = f_X(1/y)/y^2$	$F_Y(y) = 1 - F_X(1/y)$	$\mu_Y(r,\alpha_1,\alpha_2,\lambda,\sigma)$ $= \mu_X(r,\alpha_1,\alpha_2,\lambda,-\sigma)$ $= \mu_X(r,\alpha_2,\alpha_1,\lambda,\sigma)$
Generalized F (GF4) $\theta = (\alpha_1,\alpha_2,\lambda,\sigma) > 0$ $p = 1/\sigma$, $\lambda = \exp(-\mu)$	$\dfrac{p[\alpha_1(\lambda x)^p/\alpha_2]^{\alpha_1}}{B(\alpha_1,\alpha_2)\,x\,[1+\alpha_1(\lambda x)^p/\alpha_2]^{\alpha_1+\alpha_2}}$	$PROBF((\lambda x)^p, 2\alpha_1, 2\alpha_2)$	$\left(\dfrac{\alpha_2}{\alpha_1}\right)^{r\sigma} \dfrac{\Gamma(\alpha_1+r\sigma)\,\Gamma(\alpha_2-r\sigma)}{\lambda^r\Gamma(\alpha_1)\Gamma(\alpha_2)}$
Generalized gamma (GG3) $\theta = (\alpha,\beta,p) > 0$ GF: $\alpha_1 = \alpha$, $\alpha_2 \to \infty$, $\beta = \alpha^{-\sigma}e^{\mu}$, $p = 1/\sigma$	$\dfrac{p x^{\alpha p-1}\exp[-(x/\beta)^p]}{\Gamma(\alpha)\beta^{\alpha p}}$	$PROBGAM((x/\beta)^p, \alpha)$	$\dfrac{\Gamma(\alpha+r\sigma)}{(\lambda\alpha^\sigma)^r\Gamma(\alpha)} = \dfrac{\beta^r\Gamma(\alpha+r\sigma)}{\Gamma(\alpha)}$
Burr/Dubey (Bur3) $\theta = (\alpha,\lambda,p) > 0$ GF: $\alpha_1 = 1$, $\alpha_2 = \alpha$	$\dfrac{\lambda p(\lambda x)^{p-1}}{[1+(\lambda x)^p/\alpha]^{1+\alpha}}$	$1 - \dfrac{1}{[1+(\lambda x)^p/\alpha]^{\alpha}}$	$\dfrac{\alpha^{r\sigma}\Gamma(1+r\sigma)\,\Gamma(\alpha-r\sigma)}{\lambda^r\Gamma(\alpha)}$
Gumbel generalized logistic (Gum3) $\theta = (\alpha,\lambda,p) > 0$ GF: $\alpha_1 = \alpha_2 = \alpha$	$\dfrac{\lambda p(\lambda x)^{\alpha p-1}}{B(\alpha,\alpha)\,[1+(\lambda x)^p]^{2\alpha}}$	$PROBF((\lambda x)^p, 2\alpha, 2\alpha)$	$\dfrac{\Gamma(\alpha+r\sigma)\,\Gamma(\alpha-r\sigma)}{\lambda^r[\Gamma(\alpha)]^2}$

Probability Distribution	Probability Density Function (PDF)	Cumulative Distribution Function (CDF)	*r*th moment E(Xr)
Gamma (Gam2) $\theta = (\alpha, \beta) > 0$ GF: $\alpha_1 = \alpha$, $\alpha_2 \to \infty$, $\sigma = p = 1$, $\beta = e^{\mu}$	$\frac{x^{\alpha-1}\exp(-x/\beta)}{\Gamma(\alpha)\beta^{\alpha}}$	$PROBGAM(x/\beta, \alpha)$	$\frac{\beta^r \Gamma(\alpha + r)}{\Gamma(\alpha)}$
Lognormal (Log2) $\theta = (\mu, \sigma)$, $\sigma > 0$, μ *real* GF: $\alpha_1 \to \infty$, $\alpha_2 \to \infty$	$\frac{\exp\{-[(\log(x)-\mu)^2]/[2\sigma^2]\}}{\sqrt{2\pi}\,\sigma x}$	$PROBNORM\{[\log(x)-\mu]/\sigma\}$	$exp(r\mu+\sigma^2 r^2.2)$
Weibull (Wei2) $\theta = (\lambda, p) > 0$ GF: $\alpha_1 = 1$, $\alpha_2 \to \infty$	$\lambda p(\lambda x)^{p-1} \exp[-(\lambda x)^p]$	$1 - \exp[-(\lambda x)^p]$	$\frac{\Gamma(1 + r/p)}{\lambda^r}$
Log-logistic (Tic2) $\theta = (\lambda, p) > 0$ GF: $\alpha_1 = \alpha_2 = 1$	$\lambda p(\lambda x)^{p-1}/[1 + (\lambda x)^p]^2$	$(\lambda x)^p / [1 + (\lambda x)^p] F_Y(y)$	$\frac{\Gamma(1 + r\sigma)\,\Gamma(1 - r\sigma)}{\lambda^r}$
Exponential (Exp1) $\theta = \lambda = 1/\beta > 0$ GF: $\alpha_1 = 1$, $\alpha_2 \to \infty$	$\frac{\exp(-x/\beta)}{\beta} = \lambda \exp(-\lambda x)$	$1 - \exp(-\lambda x)$	$\beta^r = \lambda^{-r}$
$Y=$ mixture of X with a point mass at x=0. Y= 0 with probability M Y= X with probability 1-M	$f_Y(y) = (1 - M) f_X(x)$ for $x \neq 0$	$=0$ *for* $y < 0$ $=M+(1-M)F_X(x)$ *for* $y \geq 0$.	$E(Y^r)=(1-M)E(X^r)$

Notes: $\Gamma(\alpha)=\Gamma(\alpha)=\int_0^{\infty} x^{\alpha-1} e^{-x}\,dx$, $B(\alpha_1, \alpha_2) = \Gamma(\alpha_1)\Gamma(\alpha_2)/\Gamma(\alpha_1+\alpha_2)$. $\alpha_1 > 0$ and $\alpha_2 > 0$ are one-half the numerator and denominator degrees of freedom. μ and σ are location and scale parameters for $\log(F)$. If F is an F variate with $2\alpha_1$, and $2\alpha_2$ degrees of freedom, then the corresponding generalized (power transformed) variate is $GF=\exp(\mu+\sigma \log F)=e^{\mu}F^{\sigma}$. Formulae for moments are valid as long as α_1, α_2, σ, $\alpha+r\sigma$, and $\alpha-r\sigma$ are all positive. For X a generalized F or any of its special cases, an inverse random variable is obtained via $Y = 1/X$, with properties indicated by the row for Y=1/X. This table was prepared by Lawrence Myers of RTI and is based primarily on Prentice (1975) and Johnson and Kotz (1970).

•U.S. GOVERNMENT PRINTING OFFICE: 2000-550-101-40009