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

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

$$\frac{1}{s\sqrt{2p}}\exp\left[-\left(\frac{1}{2s^2}\right)(x-m)^2\right].$$

In this example, the two parameters of the PDF are the population mean ∶ = 5.0 and the population standard deviation F=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:

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

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

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

Another function often used to describe parametric distributions is the *cumulative distribution function* (CDF). The CDF is the probability that $X \# x_i$ 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 $X \# x_i$ for all values of x in the *sample*. The EDF is the sum of the relative frequencies of all sample values of $X \# x_i$. 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., $n \$ 1{,}000$) 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

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

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

n=the sample size (nine fish)

p=the probability of infection (0#p#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,

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

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

$$\hat{\beta}_i^2 \Big/ Var\hat{\beta}_i = \chi^2$$

with n-(1+p) degrees of freedom—[where n=the size of the sample from which MLE($\$_i$) 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-"% 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-"; for a 95% confidence interval, "=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$30. While 20#n<30 will often provide reasonably good MLEs, MLEs computed from samples of 10#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 $L\left(y\,|\,\hat{b}_0, \hat{b}_1, \hat{b}_2\right)$ = the likelihood associated with the full model

$L\left(y\,|\,\hat{b}_0, \hat{b}_1\right)$ = 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

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

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 (:) 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 ($:_1=:_1$) but with different sized variances ($F_1<F_2$), 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:

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

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 P^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$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.

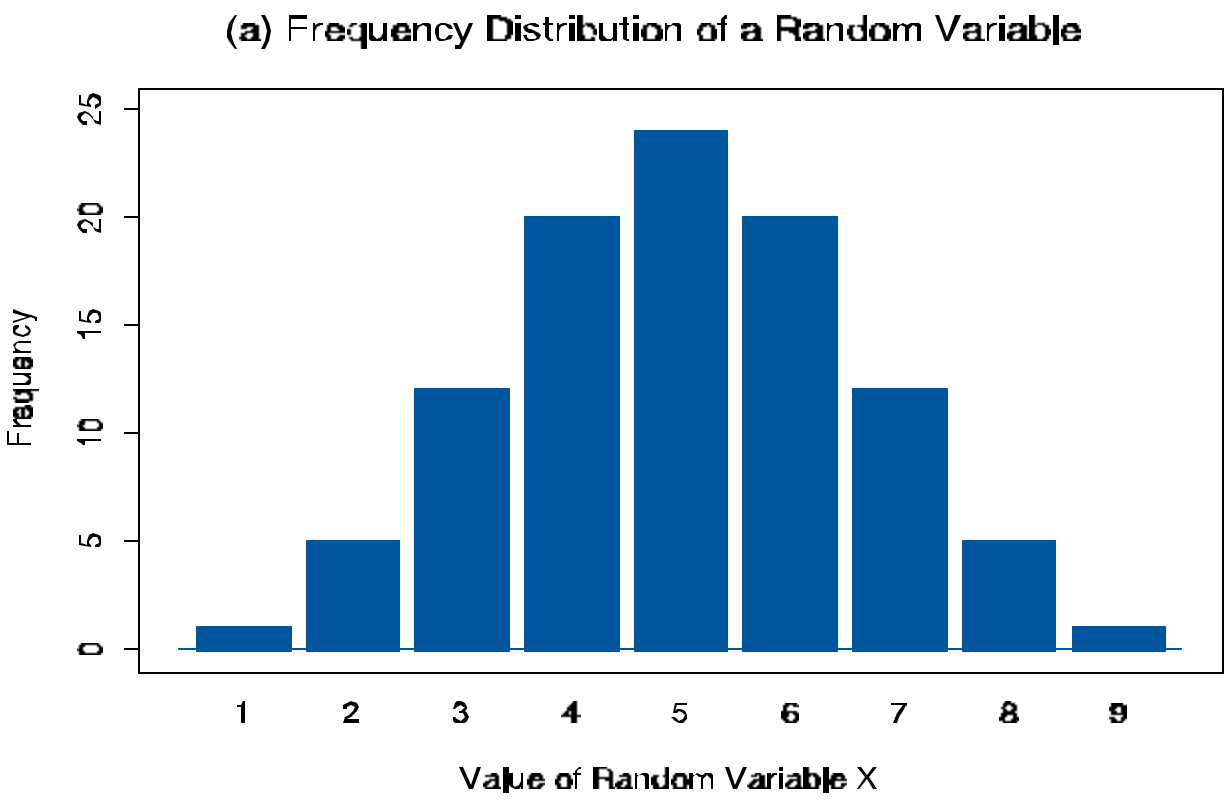
Table 1-1. Three MLEs of Prevalence, Given Different Observed Numbers of Infections

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



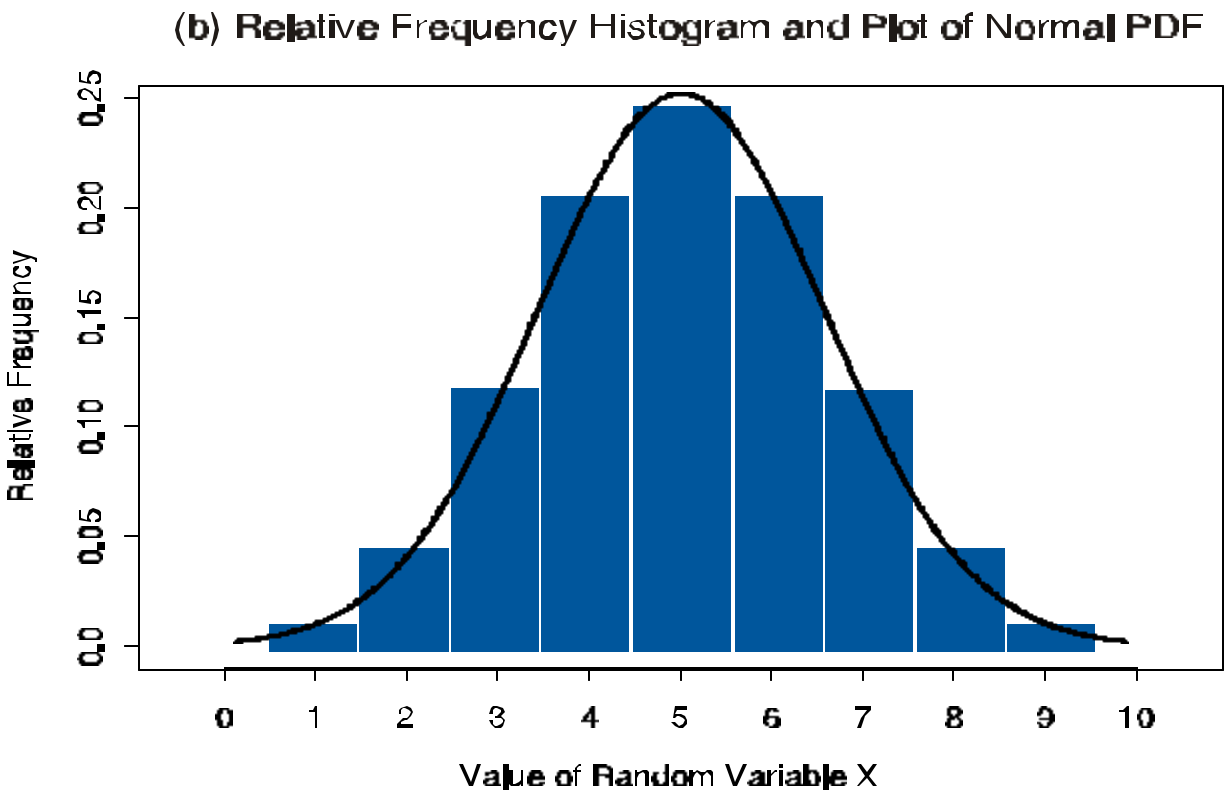
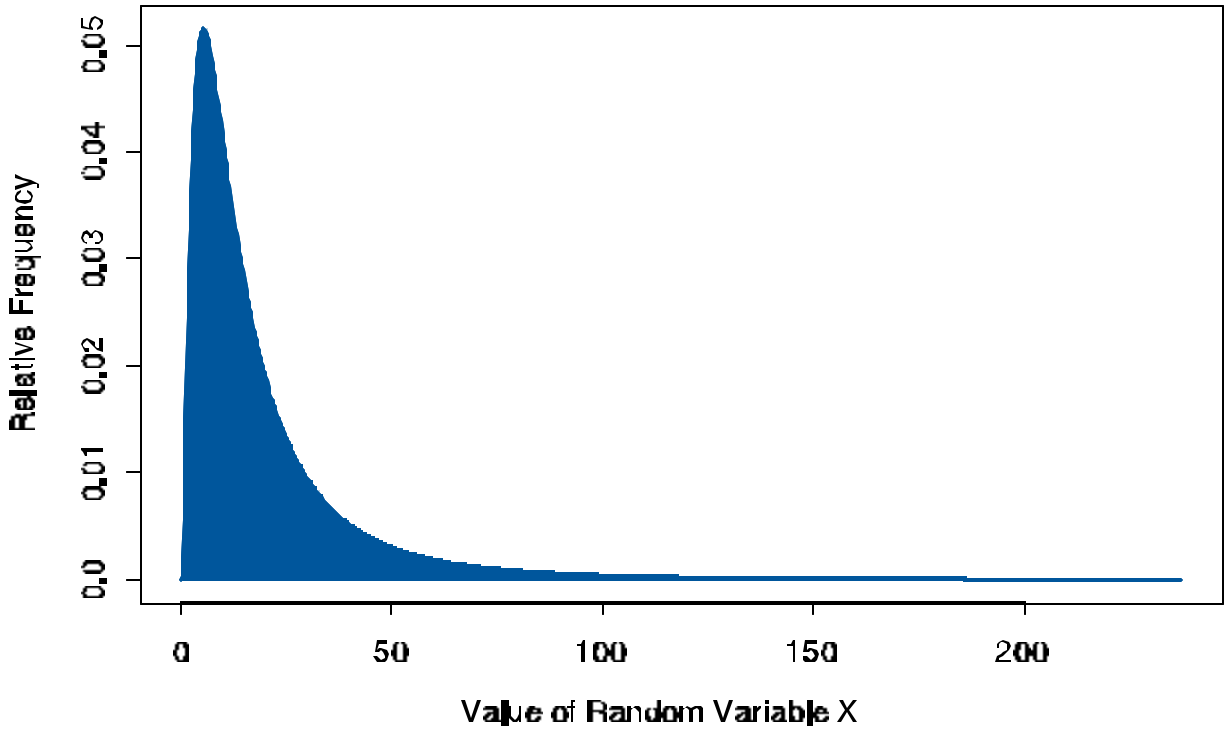


Figure 1-2. PDFs and CDFs

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



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

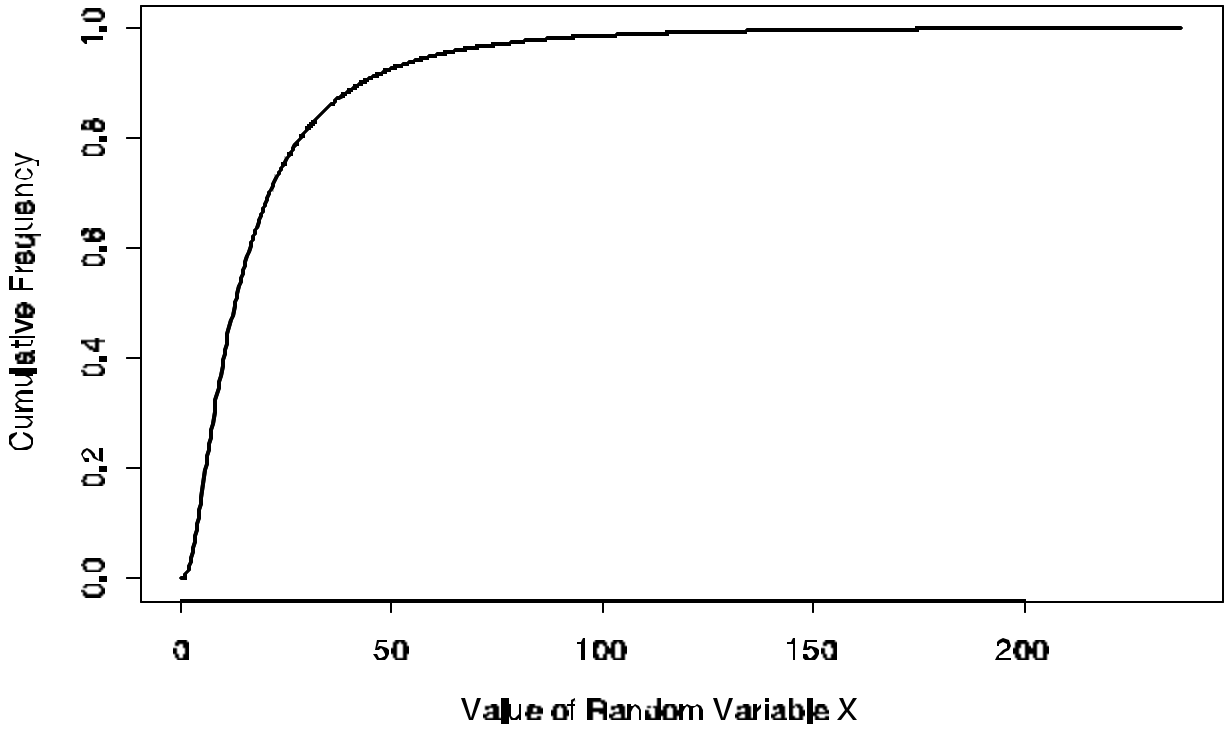


Figure 1-3. Examples of Parametric PDFs

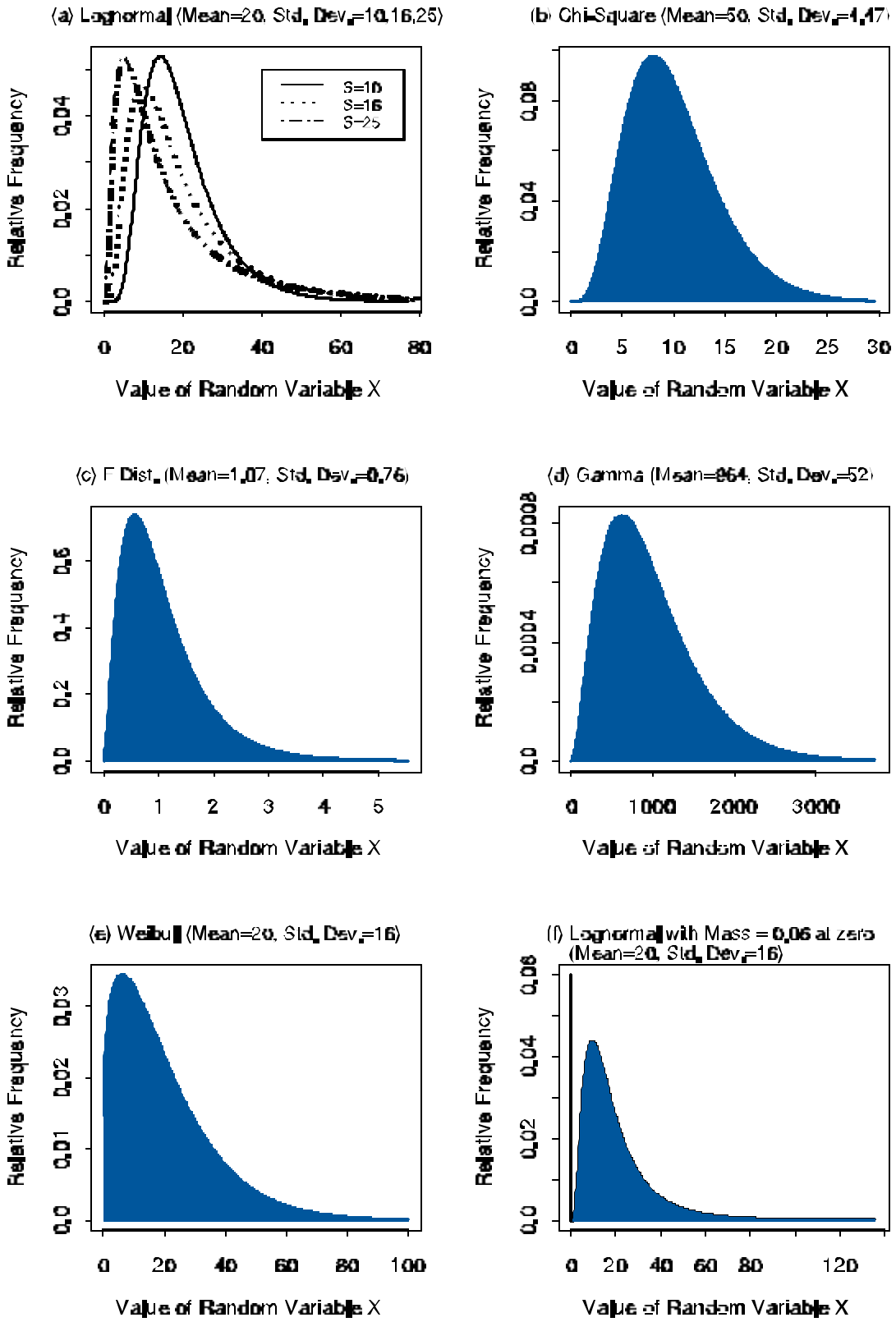


Figure 1-4. Demonstration of MLE

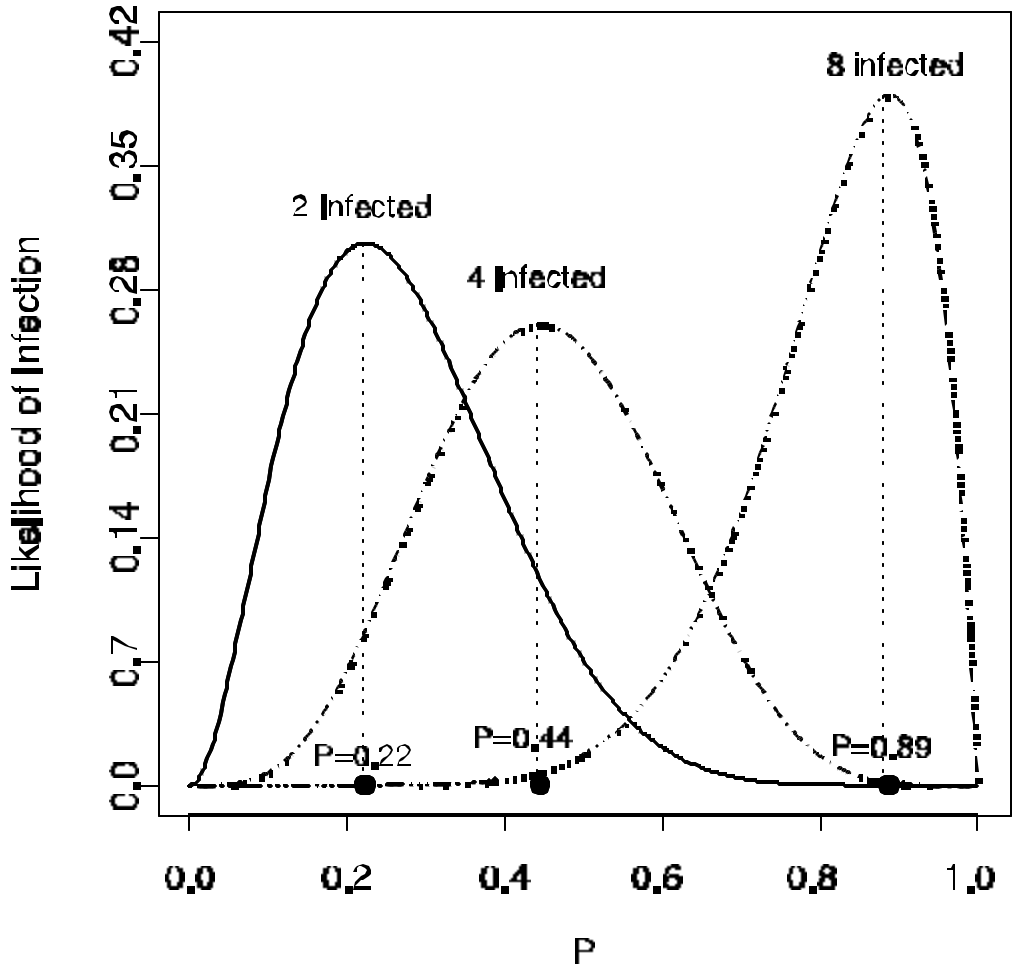
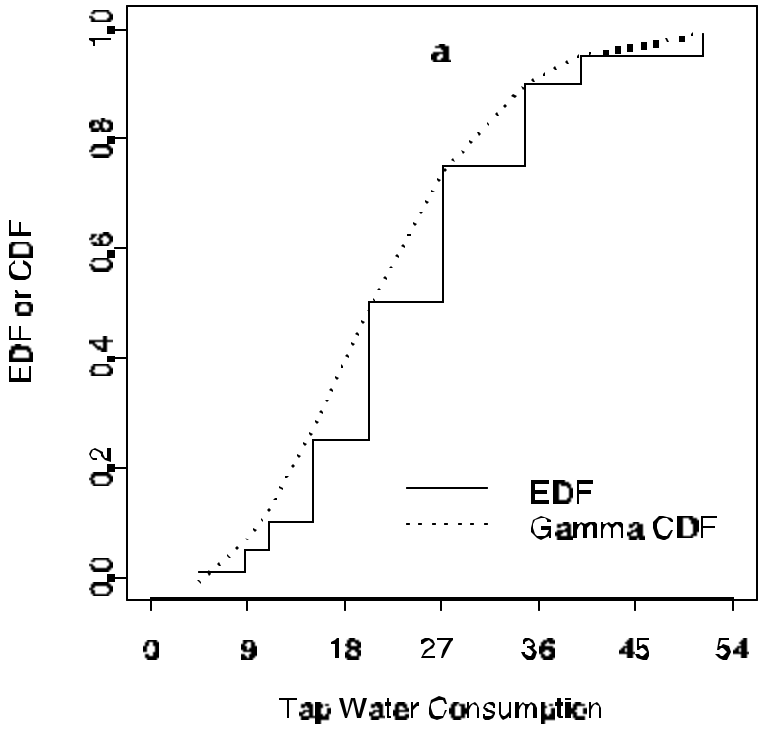


Figure 1-5. Tap Water Gamma GOF Plots: Adults 65 Years Old and Older



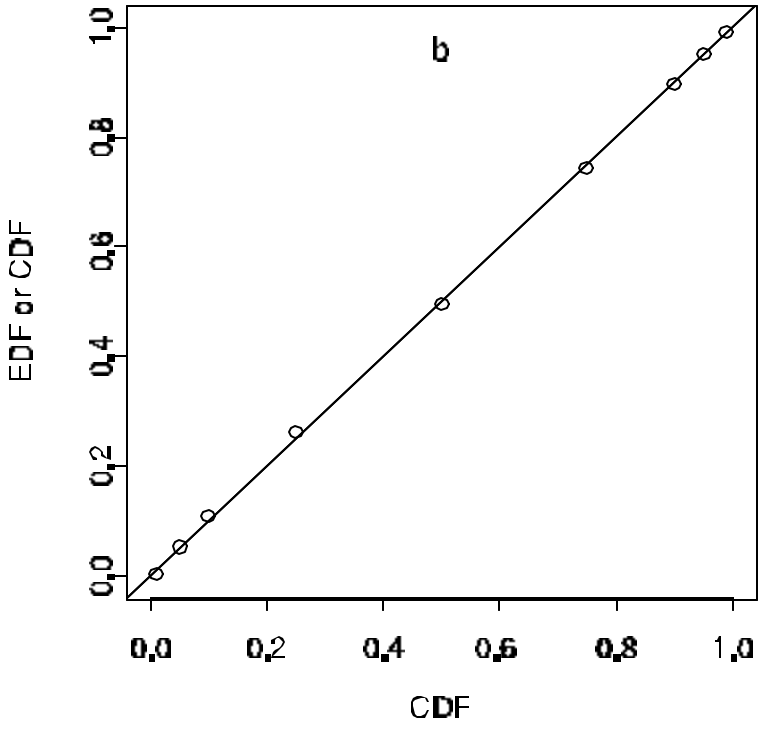


Figure 1-6. Tap Water Gamma P-P Plot and Q-Q Plot: Adults 65 Years Old and Older