Fitting Probability Distributions

1 June 2011

Prepared by

Neptune and Company, Inc.

This page is intentionally blank, aside from this statement.

CONTENTS

FIGURES

TABLES

1.0 Introduction

In the Clive DU PA model, most of the input parameters are treated as probabilistic. The term *parameter* is used to refer to any numerical quantity in the PA model. his document provides an overview of the approach to construction of probability distributions for parameters.

Note that the term parameter is used here because it is in common use in the PA and modeling community. However, since probability distributions are applied to these parameters, from a statistics perspective they should be termed variables, or even random variables.

2.0 Types of Parameters

Parameters of the PA model are mathematical constructs that represent a variety of different concepts. Assignment of a probabilistic distribution must consider the use of the parameter within the PA model.

The probabilistic behavior associated with the input may also represent a variety of different concepts. The variation may represent aleatory variability, epistemic uncertainty, or some combination of those two. The appropriate probabilistic representation for the parameter can differ greatly depending on the appropriate representation.

- Epistemic uncertainty represents lack of knowledge about the true value of the parameter. Hypothetically, data could be collected to reduce the uncertainty, which would then result in a distribution with less variation.
- Aleatory variability represents inherent randomness in the "outcome" of the parameter. The outcome may represent changes through time or space or the characteristics of individual members of a population. Given assumptions about the population or modeling assumptions underlying the parameter, further information gathering does not reduce aleatory variability. Changing the modeling or population assumptions can lead to a change in the variability (e.g. changing the spatial extent a soil porosity distribution is applied to).

Many parameters in the Clive DU PA contain at least some element of both epistemic uncertainty and aleatory variability, though the probabilistic construction is typically based on assuming one or the other. Although there are exceptions, for the most part, distributions developed assuming aleatory uncertainty are contained in the individual dose model (see the Dose Assessment white paper). Most other input distributions are developed based on epistemic uncertainty, although as noted, most parameters contain some element of both. It is often difficult to completely separate epistemic and aleatory uncertainty. Another, and perhaps better, way of framing the distinction is with respect to the spatial and/or temporal scale of each parameter. Most parameters in the Clive DU PA model represent long time frames or large areas, and the distribution of the average of the trait of interest is needed for the model. These cases are aligned more with the concept of epistemic uncertainty. However, the dose parameters are specific to individuals, representing points space, and time frames that are specific to the available data. These cases are aligned more with the concept of aleatory variability. In effect, in this model, epistemic uncertainty, upscaling and distribution of the average are related, and aleatory variability, and distribution of the data are related without the need for upscaling.

These are important distinctions in the development of complex PA models, not just for model building purposes, but also for model interpretation and comparison with performance objectives. The PA model is constructed so that raw output doses are provided for each hypothetical individual included in the model, in each year of the model. Typically, risk assessment is based on the average risk. In that context, the average dose to the individuals in each year is the relevant statistic for each receptor group (ranchers, hunters, OHV enthusiasts). Since 5,000 simulations are performed, there are 5,000 estimates of the average dose in each year of the model. If the input distributions are specified as epistemic at the appropriate spatio-temporal scale, then, by analogy with typical approaches to risk assessment, the 95th percentile of the average dose in each year is the relevant statistic of interest. This has the added advantage of properly representing uncertainty in the average dose, and hence the uncertainty can be reduced through further data collection.

Typically, doses generated from a PA are compared to performance objectives by using the "peak of the means", however, this does not adequately address the issue of dose in a year (unless the peak of the mean dose is in the same year for every simulation). There are also 5,000 estimates of the peak of the mean, however, it is not clear how to match a statistic from that distribution to the performance objectives. This model will allow exploration of this issue, to evaluate possible approaches to comparison of output doses to performance objectives.

There are other sources of uncertainty that should also be considered in a PA model. These do not fall easily into either the epistemic or aleatory categories.

- Conceptual uncertainty is typically not associated with a parameter, except in conjunction with the model as a whole.
- Numerical uncertainty is similar to model uncertainty, except that it typically relates only to the mathematical aspect of the model, and whether or not a single number can adequately represent the process.

These latter sources of uncertainty are largely ignored when constructing probabilistic distributions for parameters. These uncertainties are typically explored, to limited extent, with sensitivity analyses. However, where expert judgment is utilized in construction of a probability distribution, the presence of conceptual or numerical uncertainty may cause the expert to increase the variation associated with a parameter in order to (perhaps) produce a broader range of model outputs.

More generally, the development of distributions for model input parameters in a PA model needs to accommodate a wide range of options that address spatio-temporal scales, correlation structures, available data, secondary data, literature review information, expert opinion and abstraction from more complex sub-models. Statistical methods that can be considered in each case are described in the following sections. This is a critical component of model development. If not performed properly then the PA model runs the risk of the "garbage in – garbage out" syndrome, uncertainty and sensitivity analysis are compromised, and the results of the model are potentially meaningless. If performed properly, then everything falls into place regarding model results, comparison with performance objectives, and useful uncertainty and sensitivity analysis.

3.0 Fitting Distributions to Data

3.1 Distributions Representing Epistemic Uncertainty

When data are available, whose distribution depends on a parameter of interest, then a Bayesian approach can be used to combine any available prior information with information from the data. The posterior distribution on the parameter represents the uncertainty about the value of the parameter. Prior information could be obtained through expert elicitation, but for nearly every parameter in the Clive DU PA model for which data are available, a non-informative prior is used.

Most parameters in the Clive DU PA model correspond to physical quantities that represent an average of some type. Some parameters represent averages over time, as they represent typical behavior that will be used throughout the 10,000 year performance period, such as annual precipitation. Other parameters represent averages over space. For example, properties of vegetation represent an average vegetation effect across a model area, while soil properties represent an average across a volume of material represented by a model cell. When data are available that represent small amounts of time relative to the 10,000 years, or small areas/volumes relative to the model cells, then it is the *mean* of the data distribution that needs to be modeled. Under most regularity conditions (such as finite variance and the true parameter not on the border of the parameter space), the asymptotic distribution of a posterior distribution of a parameter is normally distributed (Gelman 2004). When a non-informative prior is used, the posterior distribution is generally well-approximated by the sample distribution of the statistic used to estimate the parameter. Thus, the posterior distribution for a mean μ is generally well-approximated by a normal distribution, according to the Central Limit Theorem, if the sample size n is sufficiently large:

$$\mu \,|\, \boldsymbol{X} \sim N\left(\bar{X}, \frac{s}{\sqrt{n}}\right) \tag{1}$$

where $\bar{X}$ is the sample mean, and s is the sample standard deviation. This approximation can be generalized to most other types of parameters, with the posterior distribution well-approximated by:

$$N\left(\hat{\theta}, s.e.(\theta)\right) \tag{2}$$

where $\hat{\theta}$ is an estimate of the parameter of interest θ, and $s.e.(\theta)$ is the standard error associated with the estimate. Stricter regularity conditions may be required for the general approximation to hold, and larger sample sizes may be needed for the posterior distribution to converge to normality.

For parameters whose sampling distributions are difficult to calculate, due to the type of parameter or the small sample size, a bootstrap approach can be utilized to simulate a sampling distribution (Efron and Tibshirani 1994). The bootstrap method simulates a sampling distribution for a parameter by simulating new sets of data of the same size and structure as the existing data.

The data simulation may be either parametric, assuming an underlying distribution for the data, or non-parametric, simulating from the empirical distribution of the data. The simulated bootstrap samples of the parameter are then fit to a distribution following the guidelines of fitting presented in Section 3.2, since the bootstrap data represent hypothetical data that can be precessed similarly to the processing of data that represent aleatory uncertainty.

3.2 Distributions Representing Aleatory Variability

For cases where the goal is to find a distribution that reflects the variability in the data, a goodness-of-fit approach is used. When the complete data set is available, the Akaike Information Criterion (AIC) is used to choose a distribution (Akaike 1974). The special case of data that are reported only as quantiles is address in Section 4.0.

AIC provides a measure of fit based on the likelihood function that attempts to discourage over-fitting by penalizing models with larger numbers of fitted parameter values. AIC could be used directly to choose a distribution by selecting the distribution that minimizes AIC. However, in order to allow for scientific judgment to choose between models that are close in fit, Akaike weights can be used for model selection (Burnham 2002). Akaike weights can be interpreted as conditional probabilities for each model when all models are treated as equally likely a priori. The Akaike approach is the following:

- Choose a set of distributions to be considered: M_1, M_2, …, M_k.
- Fit each distribution via maximum likelihood, and calculate the AIC for each model: A_1, A_2, …, A_k.
- Calculate the Akaike weights for each model:

$$W_i = \frac{e^{-0.5\cdot(A_i - A_{min})}}{\sum_{j=1}^{k} e^{-0.5\cdot(A_j - A_{min})}} \tag{3}$$

where A_{min} is the smallest AIC amongst the models being considered. Distributions with low weights are removed from consideration, and scientific considerations are used to choose between distributions with similarly high weights.

The following descriptions and figures (Figures 1 through 4) provide a list of distributions that are commonly considered for parameters in the Clive PA model: Normal, Lognormal, Gamma, Beta. Note that the uniform distribution is special case of the Beta distribution. Many other distributions are considered for special cases, but these four are adequate for most purposes. Log-uniform distributions are used for Kd and solubility as described in the Geochemistry white paper, and triangular distributions are used for a few parameters in the dose model, which represent aleatory variability, when there was insufficient data and expert elicitation has not yet been performed.

- *Normal* – N(*m*, *s*), where *m* is the mean, and *s* is the standard deviation. This distribution is unimodal and symmetric and has support on the entire real line. This distribution occurs naturally in many settings and is generally preferred for parameters representing averages or sums. Since the normal distribution has infinite support, the distribution must be left-truncated at 0 (or some other natural boundary) for certain types of parameters.

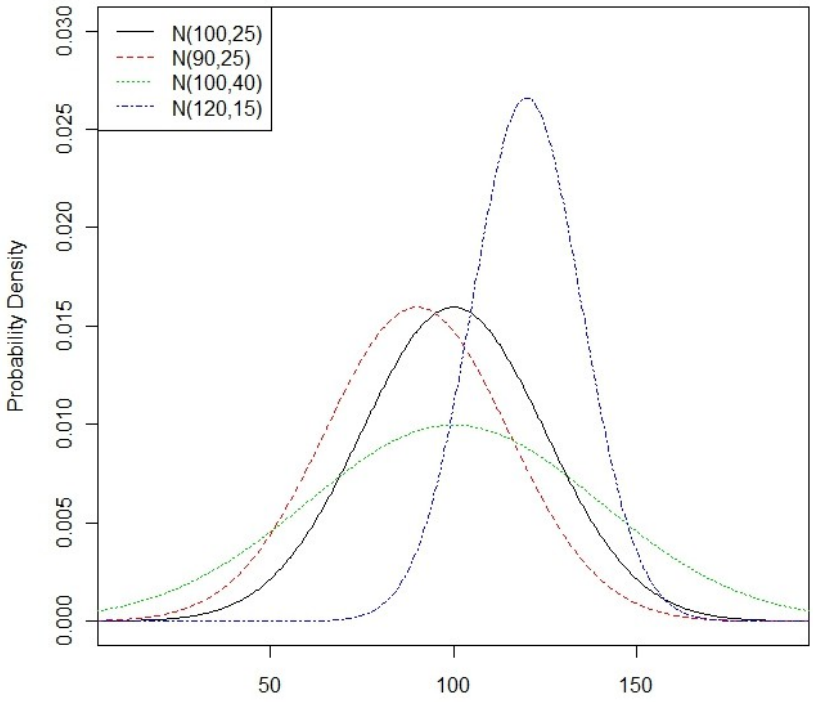

Figure 1. Examples of normal probability density functions

- *Lognormal* – LN(m, s, θ), where m is the geometric mean, and s is the geometric standard deviation, and θ is a location parameter specifying the minimum. This distribution is unimodal and right-skewed and has support on all real values greater than θ. When the geometric standard deviation is near 1, the lognormal distribution closely approximates the normal distribution. Physical quantities can often be modeled well with a lognormal distribution, and typically θ=0, forcing those quantities to be positive.

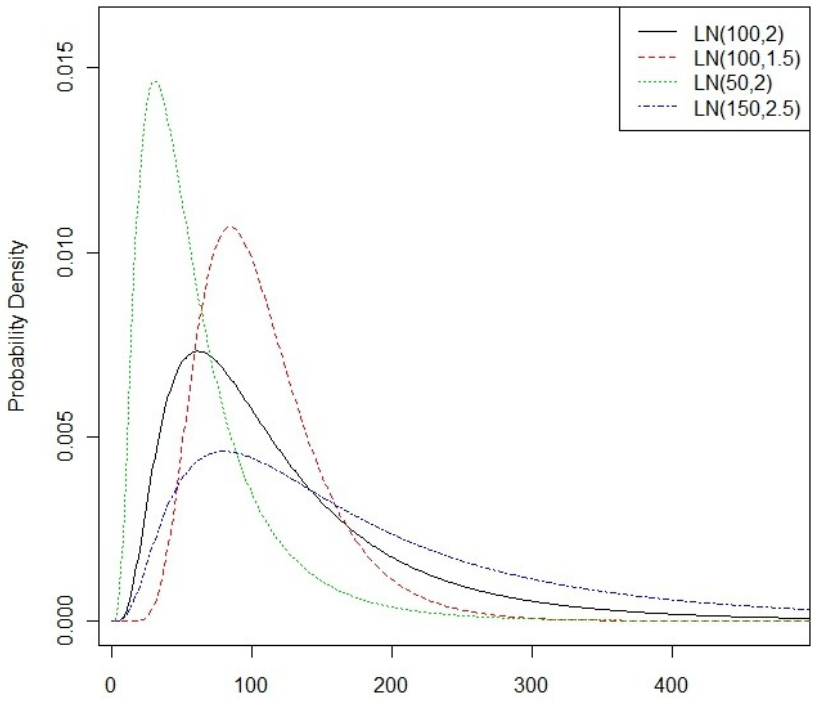

Figure 2. Examples of lognormal probability density functions

- *Gamma* – Gamma(m, s, θ), m is the mean, s is the standard deviation, and θ is a location parameter specifying the minimum. This distribution is unimodal and right-skewed and has support on all real values greater than θ. Fitted gamma distribution and lognormal distributions often appear quite similar, and the lognormal is typically preferred for physical quantities. However, the gamma distribution can fit certain types of tail behavior that the lognormal distribution cannot.

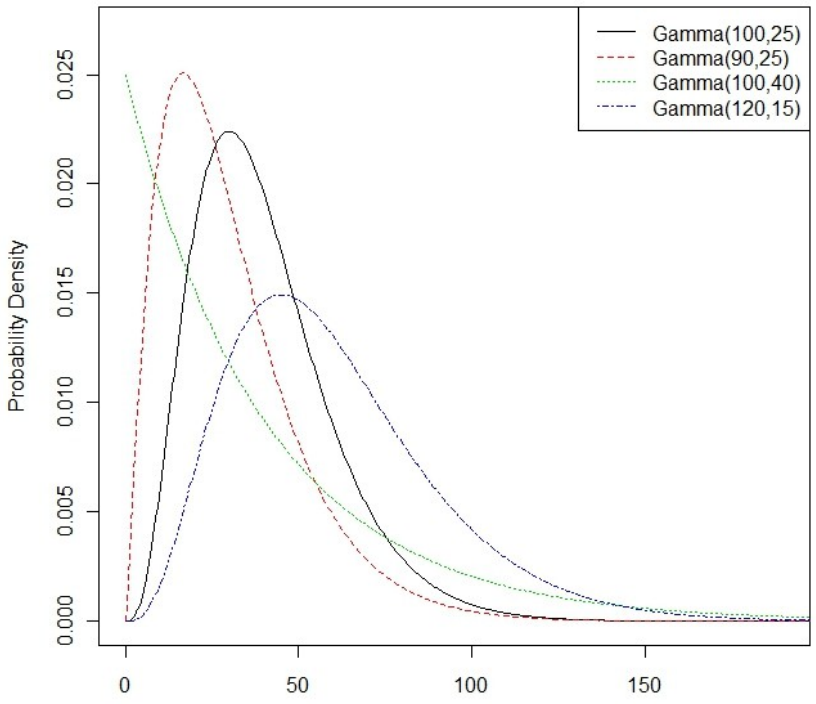

Figure 3. Examples of gamma probability density functions

- *Beta* - Beta(*m*, *s*, *l*, *u*), where *m* is the mean, *s* is the standard deviation, *l* is the lower bound, and *u* is the upper bound. The beta distribution can take on a variety of shapes. It is typically unimodal, but can be bimodal, with modes at the lower and upper bounds. The beta distribution is sufficiently flexible that it might provide a reasonable fit where other distributions cannot, and it is the only standard distribution that has finite support. For many parameters, finite support does not make good sense, so the beta distribution is typically only chosen when it is the only distribution that provides a reasonable fit, or when there is a natural lower and upper bound.

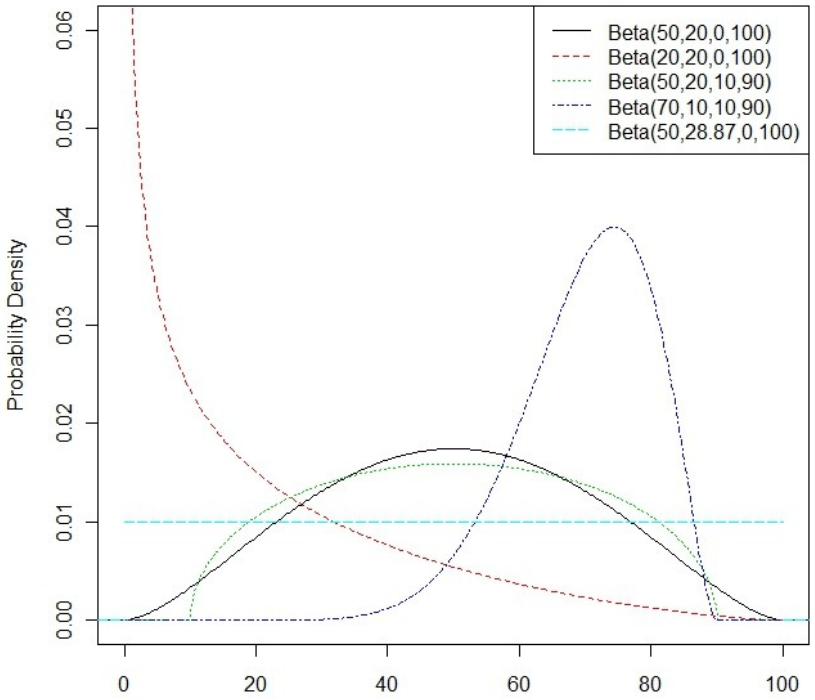

Figure 4. Examples of beta probability density functions

4.0 Fitting Distributions to Reported or Elicited Quantiles

In many cases, data are available only in the form of reported quantiles of the distribution. A formal method for fitting a distribution and choosing amongst possible distributions is needed. While the focus here is on empirical quantiles, the same approach may also apply to quantiles achieved via expert elicitation, though some assumptions about the expert's knowledge base must be considered. This section begins with a definition of quantiles, and follows up with a likelihood estimation method for estimating distributions based on quantile input, and ends with an example.

4.1 Quantiles

Let X be a random variable whose distribution is of interest. Suppose that a random sample of n observations from this distribution has been collected, $\boldsymbol{X}=\{X_i\}_{i=1}^{n}$, but that the reported summaries of this sample are restricted to a set of k empirical quantiles, $\{\hat{q}_j\}_{j=1}^{k}$, corresponding to a set of proportions $\{p_j\}_{j=1}^{k}$ (considered to be given in increasing order for convenience; i.e., $p_j < p_{j+1}$).

The empirical cumulative distribution function (CDF) is defined as:

$$\hat{F}_X(x)=\frac{\#\ of\ sample\ values\ less\ than\ \ x}{n}=\frac{\sum_{i=1}^{n} I\{X_i<x\}}{n}, \tag{4}$$

where I is the indicator function. An empirical quantile corresponds to the inverse of the empirical distribution function:

$$\hat{q}_i=\hat{F}_X^{-1}(p_i)\ . \tag{5}$$

Since $\hat{F}_X$ is a step function, the inverse is not uniquely defined. However, there are many common methods for defining a unique quantile (Hyndman and Fan, 1996). In practice, the exact method of defining the quantile is rarely cited. Thus, there is some potential error associated with a reported quantile. The relative size of the error is dependent on the underlying distribution and the quantile of interest. When sample sizes are large and/or the underlying distributions are smooth (as is the case with named families of distributions that one is likely to fit), the error associated with non-uniqueness should be small, though sensitivity analysis to this error should be performed in assessing fits based on reported quantiles. For the purposes of this document, $\hat{q}_i$ will be considered to be uniquely defined.

4.2 Likelihood Functions

If the original data set were available, then a reasonable choice for fitting the parameters of a distribution is maximum likelihood. Suppose that the random variable of interest, X, is assumed

to come from a parametric family of distributions (e.g. Gaussian, gamma, etc.), that are uniquely defined by a set of parameters $\boldsymbol{\theta}$. The likelihood function for a sample $\boldsymbol{X}$ is defined as:

$$L(\theta|\boldsymbol{X}) = \prod_{i=1}^{n} f_X(x_i|\theta) \ , \tag{6}$$

where f_X is the probability density (or mass) function corresponding to the parametric family of distributions. The maximum likelihood estimator (MLE) of the parameters is then defined by:

$$\hat{\theta} = arg\ max_{\theta} L(\theta|\boldsymbol{X}) \ , \tag{7}$$

or equivalently when maximizing the log-likelihood:

$$\hat{\theta} = arg\ max_{\theta} \ln L(\theta|\boldsymbol{X}) = arg\ max_{\theta} l(\theta|\boldsymbol{X}) \ . \tag{8}$$

When the sample has been summarized by quantiles, the likelihood function for the data takes a different form. The reported data are effectively $\boldsymbol{Y} = \{Y_j\}_{j=1}^{k+1}$, where Y_j is the number of observations between q_{j-1} and q_j.

$$Y_j = \sum_{i=1}^{n} I\{q_{j-1} < X_i \le q_j\} \ , \tag{9}$$

where $q_0 = -\infty$ and $q_{k+1} = \infty$ for notational convenience.

The reported data thus follow a multinomial distribution:

$$\boldsymbol{Y} \sim Multinomial_{k+1}(n, \boldsymbol{\pi}(\theta)) \ , \tag{10}$$

where

$$\pi_j(\theta) = F_X(q_j|\theta) - F_X(q_{j-1}|\theta) \ , \tag{11}$$

and F_X represents the CDF for X.

The likelihood function associated with the reported data is then:

$$L(\theta|\boldsymbol{Y}) = n! \prod_{j=1}^{k+1} \frac{[\pi_j(\theta)]^{Y_j}}{Y_j!} \propto \prod_{j=1}^{k+1} [\pi_j(\theta)]^{Y_j} \ , \tag{12}$$

Where proportionality is with respect to the parameters of interest, $\boldsymbol{\theta}$. Maximizing the log-likelihood is thus equivalent to maximizing:

$$l^*(\boldsymbol{\theta}|\boldsymbol{Y}) = \sum_{j=1}^{k+1} Y_j \ln\left[\pi_j(\boldsymbol{\theta})\right] = \sum_{j=1}^{k+1} n\hat{\pi}_j \ln\left[\pi_j(\boldsymbol{\theta})\right] \propto \sum_{j=1}^{k+1} \hat{\pi}_j \ln\left[\pi_j(\boldsymbol{\theta})\right] \quad , \tag{13}$$

where

$$\hat{\pi}_j = \frac{Y_j}{n} \quad . \tag{14}$$

Note that maximizing Equation (13) does not depend on knowing the sample size n, which may not be available for some data reports, and is only an abstract concept if the quantiles represent elicited values.

For most parametric families, $\pi_j(\boldsymbol{\theta})$ does not have a functional form that lends itself to analytical maximization of Equation (13). However, the CDF for most parametric families is sufficiently smooth that maximization routines work robustly.

Note also that the use of maximum likelihood estimation is similar to intent to using Bayesian statistical methods with some types of non-informative prior distributions. This approach, therefore, is similar in intent for quantile data as the methods described in Section 3.1. Use of least squares minimization instead is not recommended, because the underlying assumptions will probably not be met (e.g., normality, independence, identically distributed data).

4.3 Example: Gaussian Distribution

Suppose that data are reported as in Table 1:

Table 1. Example data, reported only as quantiles

$p_1 = 0.05 = 5\%$	$p_2 = 0.25 = 25\%$	$p_3 = 0.5 = 50\%$	$p_2 = 0.75 = 75\%$	$p_5 = 0.95 = 95\%$
$q_1 = 31$	$q_2 = 58$	$q_3 = 76$	$q_4 = 89$	$q_5 = 120$

Five quantiles are reported, and thus the data are separated into 6 bins. In fitting a Gaussian distribution to these quantiles, $\boldsymbol{\pi}$ can be expressed in terms of the standard Gaussian CDF, Φ, as in Table 2.

Table 2. Calculation of quantities for the log-likelihood

$\hat{\pi}_1 = 0.05 - 0 = 0.05$	$\pi_1 = \Phi\left(\frac{31-\mu}{\sigma}\right)$
$\hat{\pi}_2 = 0.25 - 0.05 = 0.2$	$\pi_2 = \Phi\left(\frac{58-\mu}{\sigma}\right) - \Phi\left(\frac{31-\mu}{\sigma}\right)$
$\hat{\pi}_3 = 0.50 - 0.25 = 0.25$	$\pi_3 = \Phi\left(\frac{76-\mu}{\sigma}\right) - \Phi\left(\frac{58-\mu}{\sigma}\right)$
$\hat{\pi}_4 = 0.75 - 0.50 = 0.25$	$\pi_4 = \Phi\left(\frac{89-\mu}{\sigma}\right) - \Phi\left(\frac{76-\mu}{\sigma}\right)$

$\hat{\pi}_5 = 0.95 - 0.75 = 0.2$	$\pi_5 = \Phi\left(\frac{120-\mu}{\sigma}\right) - \Phi\left(\frac{89-\mu}{\sigma}\right)$
$\hat{\pi}_6 = 1 - 0.95 = 0.05$	$\pi_6 = 1 - \Phi\left(\frac{120-\mu}{\sigma}\right)$

Maximum likelihood estimators can thus be calculated: $\hat{\mu} = 74.6$ and $\hat{\sigma} = 25.8$, resulting in a value of -1.65 for the (right portion of) Equation (13). The CDF and probability density function (pdf) for the fitted distribution are plotted in Figure 5.

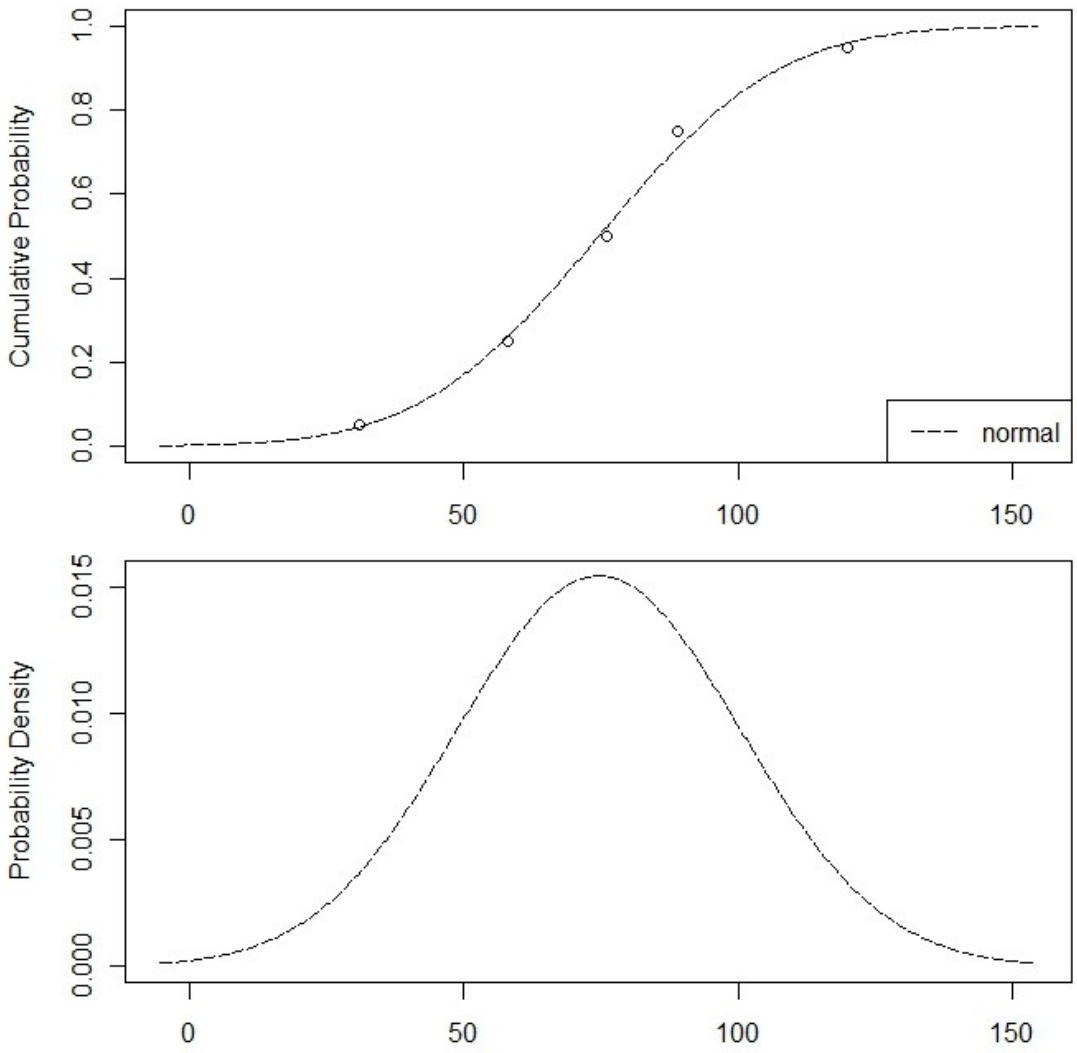

Figure 5. Fitted distribution to the quantiles of the example data

5.0 Parameter Relationships and Conditioning

Many parameters in the Clive DU PA model are related to one another. One parameter may be physically constrained by the value of another parameter, or they may simply tend to vary together. Information about the joint behavior is often unavailable, but where it is, the preferred approach is to construct joint distributions for the parameters.

When joint data are available, a simple approach is to simply calculate the sample correlation of the parameters in the data and apply the same correlation to the parameters in the model to induce a joint distribution. A simple correlation structure may not fully capture the relationship between two parameters but often provides a reasonable first approximation. Where a correlation structure is used in the Clive DU PA model, the correlation algorithms implemented in GoldSim for Gaussian copula are used (Iman and Conover 1982, Embrechts et al. 2001).

Where data and expertise are available, it is generally preferable to construct joint distributions for the parameters by constructing a marginal distribution for one parameter and *conditional* distributions for the remaining parameters. By fitting a distinct conditional distribution of the second parameter for each possible value of the first parameter, a more realistic relationship might be constructed than can be achieved through simple correlation.

For example, for the population of American males the distribution of body weight changes as a function of age, even after reaching adulthood. Beyond age 20, the median body weight tends to increase as a function of age, until middle-age, after which median body weight decreases. The variation in body weight across the population also changes with the mean. Thus, a reasonable approach might be to model body weight as:

$$BW_{males} \sim LN\left(e^{a+b\cdot Age+c\cdot Age^2}, e^{\sigma}\right) \quad . \tag{15}$$

where a, b, c, and σ are estimated from data. This general approach was utilized for the Clive PA model (including for this body weight example), by using the fitting techniques outlined in Section 4.0 to quantile data available for age and body weight.

6.0 Summary

For the Clive DU PA considerable effort has been expended to provide statistical rigor and defense for the PA model. There are few, if any, previous examples of PA for low-level waste that have achieved this level of statistical support. Regulations and guidance that could be used are sadly lacking in sufficient definition of how PA models should be constructed and the role that statistics should play to ensure proper construction. The Clive DU PA model provides an opportunity for others who perform PA for low level radioactive waste to follow this path, and improve the statistical defense for PA more generally.

7.0 References

Akaike, H. (1974). "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control* **19** (6): 716-723.

Burnham, K.P., Anderson, D.R. (2002). "Understanding AIC and BIC in Model Selection." *Sociological Methods and Research. Sociological Methods and Research,* **33** (2): 261-304.

Efron, B. and Tibshirani, R.J. (1994). *Introduction to the Bootstrap*. CRC Press LLC, Boca Raton, FL.

Embrechts, P., Lindskog, F., and McNeil, A. (2001). *Modelling Dependence with Copulas and Applications to Risk Management*, Department of Mathematics, Swiss Federal Institute of Technology, Zurich.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004). *Bayesian Data Analysis, 2nd Edition*. Chapman and Hall/CRC, Boca Raton, FL.

Hyndman, R.J., and Fan, Y. (1996). "Sample Quantiles in Statistical Packages," *American Statistician*, 50: 361-365.

Iman, R.L., and Conover, W.J. (1982). "A Distribution-Free Approach to Inducing Rank Correlation Among Input Variables," *Communications in Statistics: Simulation and Computation,*11 (3): 311-334.