

**Appendix 6:  
Software**

## Appendix 6: Software

This section contains a description of some of the software routines used in preparing the *Listeria monocytogenes* risk assessment.

### The Dose Frequency Curve-Fitting Program

This routine takes a data set that contains historical records of an association between a continuous measure and the frequency of occurrence of a discrete event in a population. It is similar in operation to ParamFit, except that it is designed to be used with data sets that correlate a dose with an outcome, rather than a simple distribution. After fitting one of more models to the data, the parameters are written to a file that may be used with the DoseFrequency object. The parameter file may also be examined with the DoseFrequency plotting routine, or the parameter estimates may be examined in Excel and used without the object.

#### *Data*

In order to proceed, the routine must be supplied with data in the proper format. There are two ways to do that.

The first is to supply a data file that is in the correct format. The “File Open” button may be used to browse for the file name. The file is not actually opened unless the “Data Edit” or “Run” buttons are selected. Alternatively, the file name, including the path, may be entered into the text box to the right of the “File Open” button.

Alternatively, data may be entered using the “Data Editor,” which is started with the “Data Edit” button. If a file name has already been entered the Data Editor will open this data file. If it has not, then the “Editor” begins with no entries.

#### *Choosing Models*

The models used by the DoseFrequency curve-fitting routine have 1 to 3 components. The total number of models fit will be equal to the number of possible permutations of each of the three components selected.

The mandatory component is the primary dose-response function listed in the “Models-to-be-Fit” box. At least one model must be selected for the program to proceed. However, any combination may be selected. The curve-fitting routine will attempt to fit all models selected. A description of the models currently supported is given below.

Models currently supported by the DF curve-fitting program and object:

Model Name	Parameters	Equation for Frequency Given Dose
Beta Poisson	alpha, beta	$1 - ((1 + (\text{dose} / \text{beta}) \text{alpha}))^{-\text{alpha}}$
Logistic	alpha, beta	$e^{\text{alpha} + \text{beta} * \ln(\text{dose})} / (1 + e^{\text{alpha} + \text{beta} * \ln(\text{dose})})$
Exponential	slope	$1 - e^{-\text{dose} * \text{slope}}$
Gompertz – Log	alpha, beta	$1 - e^{-e^{\text{alpha} + (\text{beta} * \ln(\text{dose}))}}$
Gompertz – Power	alpha, beta, power	$1 - e^{-e^{\text{alpha} + (\text{beta} * (\text{dose}^{\text{power}}))}}$
Probit	alpha, beta	$\text{normal\_cdf}(\text{alpha} + \text{beta} * \ln(\text{dose}))$
Multihit	gamma, k	$\text{gamma\_cdf}(\text{gamma} * \text{dose}, k)$
GammaWeibull	alpha, beta, gamma	$1 - (1 + (\text{dose}^{\text{gamma}/\text{beta}})^{\text{alpha}})^{-\text{alpha}}$

Ln = natural log. cdf = cumulative distribution function. ^ = raised to the power of

### *Background Parameter*

A parameter may be added to the model to accommodate other influences on the outcome of the causal event. There are three options:

- No background parameter.
- Background Dose. A background dose specified by an extra model parameter is added to the nominal dose when predicting frequency.
- Background Frequency. A background frequency specified by an extra model parameter is added to the predicted frequency. The dose-frequency function is applied to the fraction of the population who would otherwise not respond.

The program will attempt to fit all options that are checked. For example, if all three boxes are checked, then the program will examine all three different options. At least one box must be checked.

If there are fewer than five data points, then a Background Parameter cannot be employed.

### *Threshold Parameter*

If a threshold dose parameter is included in the model, and if the nominal dose is less than the threshold dose value, then the effective dose used to predict frequency is zero. If the nominal dose is greater than the threshold dose value, then the effective dose used to predict frequency is the nominal dose minus the threshold dose. The program attempts to fit all options that are checked. If both boxes are checked, then the program examines both options. At least one box must be checked. If there are fewer than five data points, then a “Threshold Parameter” cannot be employed.

### *Options*

Selecting the “Options” button opens another dialog, which gives the user some additional choices regarding how the routine operates. These include choosing the goodness-of-fit measure, how the program weights models when creating a probability tree, and the initial estimates for each of the model parameters.

### *Bootstraps*

In order to represent uncertainty arising from sampling error, dose measurement error, or the size of the exposed population in which illnesses are observed, multiple bootstraps may be performed. Sampling error, where the small sample of observed values is presumed to come from a much larger sample that is of interest, is represented by presuming a binomial distribution where the total set of values is infinitely large. The likelihood of a series of possible values for the actual frequency are computed by comparing the relative likelihood of generating the observed value. Dose and population size measurement error are sampled from distributions supplied with the data set. The total number of models fit equals the number of models selected times the number of bootstraps. While very large numbers are possible, the program has not been tested with more than 10,000 models.

### *Initial Parameter Estimates*

The default initial parameter estimates that are used by the IMSL nonlinear regression program to produce an optimum fit may not work well for all data sets. The initial estimate for the primary functions can be changed in this dialog. Initial estimates for the threshold and background parameters cannot be changed at this time. Selecting “OK” results in retention of the new parameter estimate. Selecting “Cancel” will not.

### *Model Weighting*

Even if bootstrapping is selected, the first bootstrap (the first set of models in the parameter file) always uses the original data.

### *Run*

The routine begins by fitting curves when the “Run” button is selected. The “Parameter File” dialog appears when the routine is finished. A progress bar displays the percentage of the task that has been completed. However, unless bootstrapping is selected, the results are nearly instantaneous. Selecting the “Cancel” button causes the program to exit.

As the program fits the alternative models to the data set, it calculates a weight for each of the models that is used by the object to assign probabilities to each of the models. The weighting algorithm used by the program rewards models for goodness of fit, and penalizes for parameters. Moving the slider bar to the left increases the importance of producing a good fit, while moving it to the right emphasizes the use of fewer parameters.

When bootstraps are run, the weights are recalculated on a relative basis for each bootstrap, so that the total weight for each bootstrap is identical. This means if the routine is used solely to represent the uncertainty in the parameters for a given model (*i*.

*e.*, parameter uncertainty with no model uncertainty), then the model weighting algorithm has no effect.

### **ParamFit**

ParamFit is a procedure for fitting a statistical distribution to a set of individual values for use in a subsequent Monte-Carlo simulation. It is similar in function to the routine included with Crystal Ball (Decisioneering) or BestFit, the add-on sold by Palisade as a companion to @Risk. The principle difference is that ParamFit is specifically intended for use in a population modeling exercise (*e. g.*, public health). In this circumstance, the primary purpose of a distribution is to represent variability in the measured quantity among individuals in a population, rather than the uncertainty associated with the prediction of a single event. Under such circumstances, the uncertainty is associated with the distribution used to generalize the data and draw inferences about the population as a whole. The end product of ParamFit is an Excel function containing a list of plausible alternative models which may be used to draw an inference in a two-dimensional Monte-Carlo model. It was written in Excel Visual Basic for Applications and requires Excel 5.0 or later versions and the Toxfunct add-in.

### **Models and Common Names of Their Parameters**

There are ten distributional models that may be employed for the purpose of describing the data and drawing inferences. Any models that are checked will be fit to the data. You may select only one model, all the models, or any subset. The following distributions (and the common names of their parameters) are supported by ParamFit.

#### *Statistical Distributions – Common Names of the Parameters:*

Beta: Parameter 1 = alpha, Parameter 2 = beta, Parameter 3 = minimum,  
Parameter 4 = maximum.

Cauchy: Parameter 1 = median, Parameter 2 = beta.

Exponential: Parameter 1 = lambda.

Gamma: Parameter 1 = alpha, Parameter 2 = beta.

Logistic: Parameter 1 = mean, Parameter 2 = beta.

Lognormal: Parameter 1 = geometric mean, Parameter 2 = geometric standard deviation.

Normal: Parameter 1 = mean, Parameter 2 = standard deviation.

Rectangular function: Parameter 1 = minimum, Parameter 2 = maximum.

Triangular Function: Parameter 1 = minimum, Parameter 2 = most frequent,  
Parameter 3 = maximum.

Weibull: Parameter 1 = alpha, Parameter 2 = beta.

#### *Model Weighting Criteria*

The frequency of use of each model is allocated according to its relative model weight which is calculated as follows:

$$\text{Model Weight} = (((1 + n / Pn) ^ O) * ((1 - \text{gof}) ^ H)$$

Where:

n = number of observations. This value reflects the number of data points to which the curve is fit. Values below the limit of detection (i.e. text values) are not counted for this purpose.

\* = Multiplication symbol.

Pn = Number of Model Parameters. In general, this refers to the number of parameters which are adjusted to fit the curve. However, the minimum values for the beta, linear, and triangular models would fit equally well if they are not truncated at a minimum value; that is, two points describe a line which could be represented as a single parameter (a slope).

gof = Goodness-of-Fit. ParamFit uses a least residual squares for the predicted percentiles as an optimization criteria. The ratio of the sum of residual squares to the sum of total squares for the predicted percentile is used as a goodness-of-fit statistic. This criteria emphasizes fit in the middle of the distribution, so that outliers have less impact on the shape of the distribution, other methods such as the “likelihood ratio” rate residual deviations in predicted distribution values.

O = The Parameter penalty, an arbitrary constant named after William of Ockam. Increasing this number increases the penalty for using an extra parameter, which influences the extent to which models are penalized for using an extra parameter. The maximum value is 10. Setting this value to 0 nullifies the parameter penalty.

H = The Association factor, an arbitrary constant named after David Hume. This value can be modified to increase or decrease the reward for providing a better fit. The minimum value is 0, in which case the models are weighted without regard to how well they fit the data. Increasing this value places greater emphasis on model fit.

The specific optimization criteria for the *L. monocytogenes* concentration were:

$$\text{Gof} = \sum (\text{predicted} - \text{Observed})^2 * n * \text{concentration}^{0.25}$$

Where the parameters for the goodness of fit were Predicted and Observed are the cumulative percentiles for a given concentration of *L. monocytogenes*, n is the number of samples in the report, concentration is in cfu/g. The model weight for the *L. monocytogenes* concentration =  $1 / (pN * \text{Gof}^2)$ . Where pN is the number of adjustable parameters in the distribution being fitted.