

Babar user guide

Babar v2.1

Facilia AB
May 31, 2013

Contents

1	Introduction	4
1.1	Background	4
1.2	Obtaining and installing the software	4
1.3	Structure of the document	4
2	The methods.....	5
2.1	Distributional assumptions	5
2.2	Methods for combining means and variances	6
2.2.1	Combined mean and variances.....	6
2.2.2	Pooled means and variances.....	6
2.3	Bayesian updating	7
2.3.1	Bayesian updating using conjugate priors.....	7
2.3.2	Bayesian updating using semi-conjugate priors.....	9
2.3.3	Hierarchical updating	10
2.3.4	Bayesian updating of regression coefficients.....	13
2.3.5	Convergence checking of Bayesian simulations.....	13
2.4	Weighted resampling.....	14
2.5	Distribution fitting.....	15
2.5.1	Maximum likelihood estimation (MLE)	15
2.5.2	Fitting values below detection limit.....	15
2.6	Tests of mean and variances.....	16
2.6.1	Tests of means.....	16
2.6.2	Tests of variances.....	16
3	User interface – an overview	18
3.1	The Data Editor perspective	18
3.2	The Analysis perspective.....	18
3.3	The Distribution Fitting perspective.....	19
3.4	The toolbar	20
3.5	The menus	20
3.5.1	The file menu	20
3.5.2	The Edit menu.....	20
3.5.3	The Window menu.....	20
3.5.4	The Help menu.....	20
4	Creating and managing data sheets.....	21
4.1	The project view.....	21
4.2	Adding data to the project	21
4.3	Importing a data sheet from excel	22
4.4	Export data to excel	22
4.5	Export project to excel.....	23
4.6	Editing a data sheet.....	23
4.6.1	Column Settings.....	23
4.6.2	Conversions.....	25
4.6.3	Units and unit conversions.....	26
4.6.4	Process stages.....	27
5	Performing computations	28
5.1	The Analysis data view	28
5.1.1	Switching between the current data or result of computations	28
5.1.2	The Filter tab.....	29
5.2	The Analysis tab.....	30
5.2.1	Test mean/variances.....	30

5.2.2	Pooling	31
5.2.3	Resampling.....	31
5.2.4	Direct updating.....	31
5.2.5	Hierarchical updating	32
5.2.6	Regression updating	33
5.3	Reviewing results from computations	33
5.3.1	The Analysis view.....	33
5.3.2	The Analysis Result Chart View.....	34
5.4	Inspecting results and convergence diagnostics from Bayesian simulations.	35
5.4.1	Simulation Output Statistics table view	35
5.4.2	Simulation Output Chart View.....	36
5.4.3	Simulation Information View	38
6	The Settings window.....	39
6.1.1	Application Settings	39
6.1.2	The project properties	40
6.1.3	Column Format Settings	40
6.1.4	Unit Settings and Unit Conversion Settings.....	41
6.1.5	Fitting Settings	42
6.1.6	Simulation Settings	44
7	Examples	45
7.1	Example data sheet: Nine studies of different species of bats.....	45
7.2	Example data sheet: Random measurement values.....	46
7.3	Example: Testing means and variances of species of bats	47
7.4	Example: Combining means and variances of species of bats	48
7.5	Example: Bayesian updating of a population with Daubeton's bat.....	50
7.6	Example: Hierarchical updating of eight species of bats	52
7.7	Example: Distribution fitting of observed measurements	53
7.8	Examples: Weighted resampling.....	55
8	References	57

1 Introduction

1.1 Background

Babar is an application that facilitates the derivation of probability density functions (PDFs) from measured or otherwise obtained statistics or values. The tool provides a collection of methods to test the statistical similarity of studies, to pool studies, combine studies with Bayesian updating or to fit PDFs to observed values and to data sets where some values are left-censored (e.g. below a detection limit). This document aims at providing descriptions of the methods implemented in Babar as well as the parts of the software and how to use it.

1.2 Obtaining and installing the software

Links for obtaining Babar is available at <http://www.facilia.se/projects/babar.asp>. The software will typically be installed once and can then be updated without the need to run any installer. Babar will search for updates at each startup (If the search for update feature is turned on in the application settings, see section 6.1.1). If the user confirms the update, Babar will install the updates automatically including this user guide which is available from the Help menu.

1.3 Structure of the document

Section 2 contains a theoretical explanation of the methods implemented in Babar. In many cases the formulas are accompanied with mathematical derivations or references to literature.

Section 3 contains a brief overview of the structure of the user interface.

Section 4 contains information of how to create and manage data sets, for example how to import/export data from/to excel or how to change the columns of data sheets.

Section 5 contains details of the parts of the user interface where computations are performed. The information in that section can be used as reference when following the examples in Section 7.

Section 6 provides a reference to the different settings in the settings view.

Section 7 contains examples for the methods in Babar.

2 The methods

This section aims at describing the theory underlying the methods implemented in Babar.

2.1 Distributional assumptions

Most methods assume that the data used is normally or log normally distributed. Assumptions such as these should be backed up by theoretical considerations before applying these methods.

If a variable y is normally distributed, the distributional relation is denoted

$$y \sim \text{Normal}(\mu, \sigma^2) \quad \text{Equation 2-1}$$

where μ, σ^2 is the mean and variance respectively. If a variable y' is log normally distributed, the relation is denoted

$$y' \sim \text{Log Normal}(\mu, \sigma^2)$$

or equivalently

$$\ln y' \sim \text{Normal}(\mu, \sigma^2)$$

where μ, σ^2 are the mean and variance of $\ln y$ respectively, and \ln denotes the natural logarithm. The notation y will be used to denote the normally distributed variable, or the transformation of a log normal variable. The log normal distribution is often parameterized with the geometric mean GM and standard deviation GSD. Given estimates of μ and σ^2 of $\ln(y)$ these are calculated as $GM = e^\mu$ and $GSD = e^\sigma$.

In Babar
The distribution of measurements is defined in Babar by selecting Normal or Log Normal as the value in the 'Measurement distribution' column.

The sample mean \bar{y} , sample variance s^2 and sample size n are used to summarize normally distributed samples:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

Equation 2-2

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^n (y_i - \bar{y})^2$$

If the data is instead log normally distributed the geometric mean and geometric variance are instead used. If \bar{y} and s^2 are calculated based on logarithmic data ($\ln(y)$), then the geometric mean and standard deviation are calculated as $GM = e^{\bar{y}}$ and $GSD = e^s$.

If the arithmetic mean (Mean) and arithmetic standard deviation (SD) of the untransformed log normally distributed variable is available, these can be related to the GM and GSD (Gelman, Hill 2007 pp 15):

$$GM = \frac{\text{Mean}}{\sqrt{1 + \frac{SD^2}{\text{Mean}^2}}} \quad \text{Equation 2-3}$$

$$GSD = \exp\left(\sqrt{\ln\left(1 + \frac{SD^2}{\text{Mean}^2}\right)}\right)$$

Given the $\mu = \ln(GM)$ and $\sigma = \ln GSD$, the arithmetic mean and standard deviation can be calculated as:

$$M = e^{\left(\mu + \frac{1}{2}\sigma^2\right)} \quad \text{Equation 2-4}$$

$$SD = \sqrt{(e^{\sigma^2} - 1)e^{2\mu + \sigma^2}}$$

Note: The conversion formulas between geometric and arithmetic means and standard deviations assume perfect log normality of the sample and do NOT correspond to the sample means and variances (arithmetic or geometric) calculated from the original log normal data set if it were available.

2.2 Methods for combining means and variances

The following section presents methods for combining or pooling data available from two or more studies.

2.2.1 Combined mean and variances

The method of combining mean and variances, results in a set of statistics that summarizes the data of all included studies if only the sample mean and sample variances are available. The resulting combined mean and variance are equal to the mean and variance of calculated on all data from the original data sets, if they were available known. For studies assumed to be normally distributed with sample mean μ_j and variance s_j^2 of studies $j = 1, \dots, J$ the combined mean and variance is:

$$\mu_{comb} = \frac{1}{N} \sum_j n_j \bar{y}_j$$

$$\sigma_{comb}^2 = \frac{1}{N-1} \left(\sum_j (n_j - 1) s_j^2 + \sum_j n_j (\bar{y}_j - \bar{y}_{..})^2 \right) \quad \text{Equation 2-5}$$

$$N = \sum_j n_j$$

The combined mean is weighted average of the individual means. The combined variance consists of the sum of variances within studies and the sum of variances between studies from an one way analysis of variance (ANOVA) (See section 2.6.1).

If the studies are log normally distributed, the above equation are applied to using $\bar{y} = \ln GM$ and $s^2 = (\ln GSD)^2$. The resulting statistics are then transformed back to original scale as $GM_{comb} = e^{\mu_{comb}}$ and $GSD_{comb} = e^{\sigma_{comb}}$.

In Babar

Combining means and variances are described in section 5.2.2 and example 7.4.

2.2.2 Pooled means and variances

The pooling formula takes into account the between study variation. If this variation is ignored the formula is used:

$$\mu_{pooled} = \frac{1}{N} \sum_j n_j \bar{y}_j$$

$$\sigma_{pooled}^2 = \frac{1}{N-1} \sum_j (n_j - 1) s_j^2 \quad \text{Equation 2-6}$$

$$N = \sum_j n_j$$

In Babar

Combining means and variances are described in section 5.2.2 and example 7.4.

2.3 Bayesian updating

Bayesian inference methods can be used for addressing situations where there is lack of data for the case of interest but data is available for similar cases. This is done by providing a way of combining empirical data with other available relevant information. Bayesian inference is the process of fitting a probability model to various set of data and estimating probability distributions for the parameters of the probability model. The essential characteristic of Bayesian methods is their explicit use of probability distribution for quantifying uncertainty in model parameters. This is achieved by applying Bayes' theorem which in the case of a normally distributed outcome variable is expressed as follows:

$$p(\mu, \sigma^2 | y) \propto p(y | \mu, \sigma^2) \times p(\mu, \sigma^2) \quad \text{Equation 2-7}$$

Where $p(y | \mu, \sigma^2)$ is called the data likelihood, $p(\mu, \sigma^2)$ the prior distribution of the uncertain parameters μ, σ^2 and $p(\mu, \sigma^2 | y)$ is the two dimensional posterior distribution. The relationship is proportional (\propto) since if samples can be drawn from the right hand side, it means that the correct proportion of values are draws from the left hand side. Therefore the samples can be used to draw inferences about $p(\mu, \sigma^2 | y)$. Bayes' theorem can be directly applied to estimate distribution parameters in situations where there are limited data for the "case of interest", but where other prior information is available, for example data for an analogue or a population. We wish to obtain an estimate of the distribution parameters that takes into account all information available, including prior information and new relevant data.

2.3.1 Bayesian updating using conjugate priors

For the log normal model, the fully conjugate prior distribution of μ and σ^2 is expressed in terms of the following two distribution functions (the joint two dimensional prior has been factored into two dependent prior distributions):

$$\begin{aligned} \mu | \sigma^2 &\sim \text{Normal}(\mu_0, \frac{\sigma^2}{n_0}) \\ \sigma^2 &\sim \text{Inv} - \chi^2(n_0 - 1, \sigma_0^2) \end{aligned} \quad \text{Equation 2-8}$$

where the vertical line (|) denotes that the prior of the mean, μ , is expressed using the unknown (still to be estimated) variance σ^2 . Parameters with subscript 0 are considered known and are the mean, variance and sample size (n) of the prior data set. $\text{Inv} - \chi^2(v, \sigma^2)$ denotes the Scaled Inverse Chi Square distribution with v degrees of freedom and scale parameter σ^2 . This distribution is derived from the standard $\chi^2(v)$ (Chi-Square) distribution. A sample from $\text{Inv} - \chi^2(v, \sigma^2)$ is obtained as $Y = v\sigma^2/X$ where X is a sample from $\chi^2(v)$. When combined with new data, the prior distributions (Equation 2-8) are updated and the posterior will be of the same form but with new parameters:

$$\mu|\sigma^2, y \sim \text{Normal}(\mu_n, \frac{\sigma^2}{n_0 + n}) \quad \text{Equation 2-9}$$

$$\sigma^2|y \sim \text{Inv} - \chi^2(v_n, \sigma_n^2) \quad \text{Equation 2-10}$$

Where parameters with subscript n reflect the combined prior and data:

$$\mu_n = \frac{n_0\mu_0 + n\bar{y}}{n_0 + n}$$

$$\sigma_n^2 = \frac{1}{v_n}((n_0 - 1)\sigma_0^2 + (n - 1)s^2 + \frac{n_0n}{n_0 + n}(\bar{y} - \mu_0)^2) \quad \text{Equation 2-11}$$

$$v_n = n_0 + n - 1$$

Drawing inferences from the posterior distributions of μ and σ^2

Inferences of the posterior distribution is performed by obtaining samples from the marginal posterior distributions $p(\mu|y)$ and $p(\sigma^2|y)$. This is done by sampling iteratively from the conditional posteriors, first obtaining a value from the posterior distribution of σ^2 in Equation 2-10 and then of μ in Equation 2-9 using the previously draws value of σ^2 .

In Babar, the posterior distributions are summarized with percentiles, mean and variances and a measure R (the Gelman Rubin convergence statistic) of the convergence of the obtained samples to the posterior distribution. Babar uses the medians of the obtained samples posterior distribution of μ and σ^2 to present the estimated mean and variance of the combined distributions. For log normal studies, the GM and GSD are derived as $GM = \exp \hat{\mu}$ and $GSD = \exp \hat{\sigma}$, where $\hat{\mu}$ and $\hat{\sigma}^2$ are the medians of the marginal posterior distributions. An alternative method of estimating μ and σ^2 implemented in Babar is by basing the point estimates of μ and σ^2 on the predictive distribution of y given the posterior samples. Samples from the predictive distribution is then obtained by drawing from the measurement model $y_k \sim N(\mu_k, \sigma_k^2)$ for each set of posterior samples $k = 1, \dots, K$. The mean and variance of the K obtained predictive samples are then used to estimate the μ and σ^2 respectively. This method of estimating the distribution parameters takes into account the posterior uncertainty of the distribution parameters, which the method of taking medians do not. Therefore it provides a more conservative estimate of the posterior variance. However, the method often provide unrealistic estimates of the SD/GSD (typically when the estimation is based on very few samples and/or weak priors).

The method conjugate updating is reasonable when the prior information takes the form of n_0 number of samples from a population with variance σ^2 (Gelman et al., 2004). That is, both the information from the observed data and the prior can be expressed with the sample mean, variance and number of samples and the sample variances estimates the same true population variance.

Comparing Conjugate updating with the combination of means and variances

The expressions for μ_n and σ_n^2 are equal to the expressions for the combined mean and variance (Equation 2-5) of the prior and observed data sets. This can be seen if the expression for σ_n^2 is rewritten as

$$\sigma_n^2 = \frac{1}{v_n} ((n_0 - 1)\sigma_0^2 + (n - 1)s^2 + n(\bar{y} - \mu_n)^2 + n_0(\mu_0 - \mu_n)^2)$$

Thus the method of updating with a conjugate prior produces similar results as the method of combining means and variances in 2.2.1.

In Babar
Conjugate and semi conjugate updating are described in section 5.2.4 and example 7.5.

2.3.2 Bayesian updating using semi-conjugate priors

If the prior information does not take the form of a sample with known sample size and variance σ^2 , Gelman et al, 2004 suggests the use of independent prior distributions of the mean and variance. The use of one normal prior distribution of the mean and one *Inv - χ^2* distribution of to the variance is termed semi-conjugate priors. With such priors, the *conditional* posteriors $p(\mu|\sigma^2, y)$ and $p(\sigma^2|\mu, y)$ still take the same functional form as the priors, but with updated parameters.

The method can be used when the prior information takes the form of subjective or otherwise derived normal distribution of the mean value. For the variance, an independent informative prior could be used, or assigned a so called non-informative prior distribution indicating that no prior knowledge of the variance is available beforehand.

The conditional posterior distribution of the mean becomes:

$$\mu|\sigma^2, y \sim \text{Normal}(\mu_n, \tau_n^2)$$

Equation 2-12

$$\hat{\mu}_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

$$\tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

The posterior mean is thus a weighted combination that reflects the amount of information available about the sample mean \bar{y} and the prior mean μ_0 . The amount of weighting is determined by the squared standard error σ^2/n of the data and the variance τ_0^2 of the prior distribution of the mean: A small standard error of the measurements and/or a large prior variance of the mean pulls the posterior mean closer to the sample mean.

If the variance is considered known from data then σ^2 is replaced by the sample variance s^2 in Equation 2-12. If the variance is considered uncertain or if prior information about the variance is available, the posterior distribution (conditioned on the mean μ) becomes:

$$\sigma^2|\mu, y \sim \text{Inv} - \chi^2(v_n, \sigma_n^2)$$

Equation 2-13

$$v_n = v_0 + n$$

$$\sigma_n^2 = \frac{v_0\sigma_0^2 + (n - 1)s^2 + n(\bar{y} - \mu)^2}{v_0 + n}$$

The posterior variance is expressed in terms of a weighted combination of the sample variance s^2 , an estimate of the prior variance σ_0^2 and the squared distance between the data and posterior mean. The weights are the number of measurements n and the prior degrees of freedom ν_0 for the variance respectively.

Babar supports the use of known variance (taken as the s^2) or uncertain variance with a non-informative prior.

Drawing inferences from the posterior distributions

If the variance is considered known from data, the expression for $\mu|\sigma^2, y$ in Equation 2-12 can be sampled from directly if the variance σ^2 is replaced with the sample variance s^2 .

If the variance is considered unknown or if prior information about σ^2 is to be included, then both equations Equation 2-12 and Equation 2-13 are sampled from iteratively. The method of iteritvly sampling from the full conditional distributions is called Gibbs Sampling. In each iteration a value is drawn from the conditional posterior of σ^2 from Equation 2-13 and then of μ in Equation 2-12 using the previously drawn value of σ^2 . When drawing the first sample of σ^2 a crude starting value must be used for μ , such as the sample mean (Gelman et al, 2004). Repeating this many times yields a collection of samples from the joint posterior $p(\mu, \sigma^2|y)$ and inferences can be drawn by calculating statistics of interest from the samples. Due to the arbitrary choice of start value, the first samples should not be used in inferences and should routinely be removed before inferences are drawn (called burn in sample size). The algorithm is run for different choices of start values, randomly dispersed around the maximum likelihood estimates (the number times to run the algorithm with different start values is often denoted *number of chains*). For a more thoroughly review of the Gibbs Sampling algorithm the reader is referred to literature such as / Casella and George, (1992) and Gelman et al., 2004/.

In Babar
Conjugate and semi conjugate updating are described in section 5.2.4 and example 7.5.

2.3.3 Hierarchical updating

Consider a number of related units (such as sites or species) or groups of measurements that are believed to be similar. The hierarchical model is suitable when making estimates for all quantities simultaneously, letting the units borrow strength from the ensemble (Morris, C. N., 1983). The method offers an alternative to using separate estimates and a complete pooled estimate for the units by estimating the mean value of each unit and at the same time incorporating data from all included units. The estimates from hierarchical models are therefore sometimes called partially pooled estimates or shrinkage estimators (Gelman & Hill, 2007).

In a hierarchical model with J units (e.g. J species), the mean of unit j, is modeled as coming from a common population distribution

$$\mu_j|\tau^2 \sim N(\mu, \tau^2), \quad j = 1, \dots, J \tag{Equation 2-14}$$

where the parameter μ, τ^2 are the mean and variance of the population (called hyper parameters). When the hierarchical model is fitted to data, posterior distributions are obtained for each of the unit's means μ_j , as well as for the hyper-parameters.

The conditional posterior of the mean of unit j, is (Gelman et al., 2004):

$$\mu_j | \sigma_j^2, y \sim \text{Normal}(\mu_n, \tau_n^2)$$

Equation 2-15

$$\hat{\mu}_n = \frac{\frac{1}{\tau^2} \mu + \frac{n_j}{\sigma_j^2} \bar{y}_j}{\frac{1}{\tau^2} + \frac{n_j}{\sigma_j^2}}$$

$$\tau_n^2 = \frac{1}{\frac{1}{\tau^2} + \frac{n_j}{\sigma_j^2}}$$

The amount of pooling from the unit mean \bar{y}_j to the population mean μ is determined by population variance τ^2 and the squared standard error of individual estimate, with complete pooling as special case when $\tau^2 \rightarrow 0$ and/or when the squared $\sigma_j^2/n_j \rightarrow \infty$.

In a full Bayesian treatment the hyper-parameters are assigned prior distributions to reflect some prior knowledge or belief about them. When no such information is available, so called non-informative priors are often used. This will let the hyper-parameters be estimated without introducing any explicit prior knowledge on them (Gelman et al, 2004). For non-informative priors of the hyper-parameters, the posterior distributions of the population mean μ and variance τ^2 conditioned on the parameters in the lower levels of the hierarchy, are (Gelman et al, 2004):

$$\mu | \mu_1, \dots, \mu_J, \tau^2, y \sim N(\hat{\mu}, \frac{\tau^2}{J})$$

Equation 2-16

$$\hat{\mu} = \frac{1}{J} \sum_{j=1}^J \mu_j$$

$$\tau^2 | \mu_1, \dots, \mu_J, y \sim \text{Inv} - \chi^2(J - 1, \tilde{\tau}^2)$$

$$\tilde{\tau}^2 = \frac{1}{J - 1} \sum_{j=1}^J (\mu_j - \mu)^2$$

Equation 2-17

That is, the population mean is expressed in terms of the average of the units' posterior means with variance of given by the population variance scaled by the number of units. The posterior population variance is simply the variance of the units' posterior means around the population mean.

Assumptions about within unit variances

The within unit variances can be modeled as different (heteroscedastic) or equal (homoscedastic). If variances are modeled as similar, then σ_j^2 can be replaced with a common σ^2 in Equation 2-15. If the variances are modeled as different, then σ_j^2 is replaced with the sample variance or estimated with a prior.

In the case of a common variance with a non-informative prior, the posterior becomes:

$$\sigma^2 | \mu_1, \dots, \mu_J, y \sim \text{Inv} - \chi^2(v_n, \sigma_n^2)$$

$$v_n = \sum_{j=1}^J n_j \quad \text{Equation 2-18a}$$

$$\sigma_n^2 = \frac{1}{v_n} \left(\sum_{j=1}^J [(n_j - 1)s_j^2 + n_j(\bar{y}_j - \mu_j)^2] \right)$$

The posterior variance is thus expressed as the pooled variance of all units adjusted for the updated mean values (the squared difference in the last term of σ_n^2).

If the variances are considered unequal, the posterior of the variance of unit j becomes:

$$\sigma_j^2 | \mu_1, \dots, \mu_j, y \sim \text{Inv} - \chi^2(n_j + v_0, \sigma_{n_j}^2) \quad \text{Equation 2-19b}$$

$$\sigma_{n_j}^2 = \frac{1}{n_j + v_0} \left(v_0 \sigma_0^2 + (n_j - 1)s_j^2 + n_j(\bar{y}_j - \mu_j)^2 \right)$$

A common prior is here assumed with scale σ_0^2 and degrees of freedom v_0 . A non-informative prior is obtained by setting $v_0 = 0$.

The individual posterior means of the units in the hierarchical model is partially pooled towards the population mean similar to what was achieved in method of using semi-conjugate priors in (2.3.2), in that the within unit variance, the within units sample size and the population variance determines the amount of pooling of each unit. However, in the hierarchical model the unspecified population distribution acts as the prior distribution of the units' means and is estimated instead of explicitly given.

Drawing inferences from the posterior distributions

The estimation of the joint posterior distribution of all the involved parameters requires the iterative sampling from each of the conditional posterior distributions (Equation 2-14 to

Equation 2-18a) as follows:

0. Start with crude estimates of μ_1, \dots, μ_j and μ , for example as the units' sample means and the average of the sample means.
1. Sample from $p(\tau^2 | \hat{\mu}_1, \dots, \hat{\mu}_j, \hat{\mu}, y)$ using the previously obtained sample or estimate of the units' means and population mean μ .
2. Sample from $p(\mu | \hat{\mu}_1, \dots, \hat{\mu}_j, \hat{\tau}^2, y)$ using the previously obtained sample or estimate of the units means and the population variance.
3. Sample from $p(\sigma^2 | \hat{\mu}_1, \dots, \hat{\mu}_j, y)$ using the previously obtained sample or estimate of units means.
4. Sample from $p(\mu_j | \hat{\mu}, \hat{\sigma}^2, y)$ for $j=1, \dots, J$ using the sample for σ^2 obtained in 1 and the previous sample or estimate of μ .

Step 1-4 are then repeated many times (e.g. 10 000 or 100 000) resulting in a collection of samples of all model parameters from the joint posterior distribution. To assure convergence of the samples to the true posterior distribution, it is common to run the simulation a few times with different start values (often by adding a random component to the crude estimates). To diminish the impact of the arbitrary start values, the first samples must be discarded before drawing inferences from the posterior distributions. Inferences are then drawn by calculating statistics of interest (such as the mean, median or standard deviation) of samples for the

parameter of interest. For details about the implementation of the Gibbs Sampler and issues of convergence see(Casella and George, (1992) and Gelman et al., 2004). With very few units of measurements (small J) the uncertainty in the estimated τ^2 can be large, resulting in very little pooling or even difficulties to converge. This is especially true when non-informative priors are used for the hyper-parameters. With the non-informative prior for τ^2 used here, the theoretical lower bound for the number of included units is three, but five units or less can be problematic (Gelman, 2006).

In Babar
Hierarchical updating is described in section 5.2.5 and example 7.6.

2.3.4 Bayesian updating of regression coefficients

Regression updating, extends the measurement model $y \sim N(\mu, \sigma^2)$ with a linear regression model for the mean (or log mean):

$$\mu = b_0 + b_1 * X_1 + \dots + b_k * X_k$$

The variance σ^2 then quantifies the error in the model from the observed values.

Instead of just estimating μ, σ^2 the goal is not to estimate the unknown or uncertain parameters b_0, \dots, b_k and σ^2 . The variables X_1, \dots, X_k are observed regression variables/independent variables that are presumed to have some correlation with the outcome variable y.

Bayesian regression assigns prior distributions to each of the uncertain parameters b_0, \dots, b_k Babar supports Normal distributions as prior distributions for the parameters. This is considered sufficient in many situations since the information about the coefficients is often summarized with a mean and standard error. The prior for the regression variance parameter σ^2 is assumed “non-informative” which let it be estimated from the available data.

The posterior distributions of the parameters can be expressed analytically but are interdependent for all estimated parameters and must be expressed in matrix notation. The full expressions for these are found in Gamerman and Lopez, 2006.

Drawing inferences from the posterior distributions

The posterior distributions of the k+1 uncertain parameters is summarized with statistics such as mean, standard deviations and percentiles.

2.3.5 Convergence checking of Bayesian simulations

The Gibbs sampler is a special case of a collection of algorithms called Markov Chain Monte Carlo (MCMC) methods. It is generally stated, that samples obtained with such methods must be checked for convergence before used. The reason twofold, 1) they rely on an arbitrary start value for the first iteration 2) for complicated cases the algorithm can “get stuck” for a number of iterations (especially if the posterior is multimodal). In a more general MCMC implementation called Metropolis Hastings a third complication is that values are often repeated in subsequent draws (that is, an iteration has only a certain probably of changing the value of the Markov Chain). Because of these issues, the obtained samples must be evaluated before one uses them as representative draws from the posterior distributions.

The methods currently supported in Babar does not result in multimodal posterior distributions and the Gibbs sampler accepts each draw of the conditional posterior distributions in each

iteration. The choice of random start values for semi-conjugate and hierarchical updating however still requires the convergence of the samples to be assessed.

Gelman Rubin Convergence statistic

The measure of convergence adopted here is the Gelman Rubin convergence statistic:

$$R = \sqrt{\frac{\hat{V}}{W}} \tag{Equation 2-20}$$

Here, \hat{V} is an estimate of the variance of the posterior distribution

$$\hat{V} = (n - 1)W + B/n + B/(m \cdot n) \tag{Equation 2-21}$$

Here, B/n is the variance between the means of m chains, with n values of each chain:

$$B/n = \sum_{i=1}^m (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2 / (m - 1) \tag{Equation 2-22}$$

And W is the average of the m within chain variances s_i^2 each based on $n-1$ degrees of freedom

$$W = \sum_{i=1}^m \frac{s_i^2}{m} \tag{Equation 2-23}$$

The statistic Equation 2-23 measures the *potential variance reduction* possible by obtaining more samples. It is always > 1 and a value close to 1 denote little or no potential reduction of variance. A value of $R < 1.001$ is often recommended in the literature.

All Bayesian simulations in Babar are performed in a minimum of three chains with dispersed starting values centered around an estimate from data (typically maximum likelihood estimates). For the hierarchical model, start values of the hyper parameters are estimated using the maximum likelihood estimates of the parameters on the lower level.

Monte Carlo Standard Error of the mean (MCSE)

The Monte Carlo Error of the mean (MCSE) quantifies the precision of the mean of the posterior samples. It is defined as $SD_{posterior} / \sqrt{n_{posterior}}$ where SD is the standard deviation of the posterior samples and n is the number of posterior samples. The MCSE can be interpreted as follows: If the posterior mean is 5.4321 and the MCSE is 0.01, then the posterior mean is correct to the first decimal.

In Babar
Convergence and summary statistics of posterior quantities are checked in Babar in the Simulation Output view (section 5.4.1) and the Simulation output charts (section 5.4.2).

2.4 Weighted resampling

The method of weighted sampling randomly from K probability density functions, with the proportion of samples representing each PDF given by an integer weight $n_k > 0$. The sampling is performed as follows:

Let S_{total} be the total number of samples to obtain from the sampling. To achieve the correct proportion of samples the procedure can sometimes return slightly more samples than S_{total} . The sampling procedure is defined as follows: For each PDF k ,

- 1) The proportion of values to draw from PDF k is calculated as $w_k = \frac{n_k}{\sum n_k}, k = 1, \dots, K$
- 2) The number of samples to draw from PDF k is calculated as $s_k = \text{ceil}(w_k \cdot S_{total})$ where $\text{ceil}(x)$ is the ceiling function that gives the integer that is closes to but larger than x .
- 3) s_k samples is drawn from PDF k

The method results in $N = \sum s_k$ samples that can be used to characterize the PDFs and standard distribution functions can be fitted to the obtained samples. To fulfill the proportions s_k , the actual number of samples N can be equal to or slightly larger than the wanted number of samples S_{total} .

In Babar
Weighted resampling is described in section 5.2.3 and example 7.8.

2.5 Distribution fitting

Standard Probability Density Functions (PDFs) can be fitted to measurement values or samples generated by weighted resampling. The default method of fitting distribution parameters is the maximum likelihood method. If there are values below detection limit in the fitted data set, a method taking these values into account can be used.

After fitting the distribution parameters, the Kolmogorov-Smirnov (KS) test statistics is calculated for each PDF. The KS-statistic is defined as the maximum deviation between the hypothesized cumulative distribution function and the empirical cumulative density function and is a measure of the discrepancy of the tested PDF and the data. The fitted distributions can be ranked in order of decreasing test statistic. Note that the KS test statistic is only one or other possible measures of the goodness of fit.

It is required that there are at least three observed values to fit distributions to the data.

2.5.1 Maximum likelihood estimation (MLE)

The default method of fitting distribution parameters is the maximum likelihood method. The values of the parameters of the distribution are then taken as the values that maximize the likelihood function:

$$L(\theta_1, \dots, \theta_k | y) = \prod_{i=1}^n f(y_i | \theta_1, \dots, \theta_k)$$

For some distributions the MLE parameters are analytically derived. For others the estimation is done by numerical optimization algorithms.

2.5.2 Fitting values below detection limit

If one or more of the values is only known to be below a certain value it is called left-censored. A left-censored value is specified using the "less than" sign (e.g. "<0.01"). To fully use the specified information, the following method (Burmester and Hull 1997) based on the empirical cumulative distribution of the values is used. This method is only applicable for Normal or Log normal distributions.

For the n_{obs} completely observed values, the empirical cumulative distribution is calculated. That is the following values are calculated:

$p_k = (\text{The number of values that are below } x_k) / (\text{Number of total values})$
 $z(p_k) = \text{The inverse of the cumulative standard normal distribution evaluated at } p_k.$

Note that the values that are below detection limit are used to calculate p_k for the observed values.

A regression line is fitted to the values $(x_i, z_i), i = 1, \dots, n_{obs}$. The resulting intercept and slope is taken as the mean and standard deviation respectively of the fitted distribution for all values (including the values below detection limit). A Log Normal distribution is fitted by applying the above procedure for logarithmic values.

2.6 Tests of mean and variances

Babar provides methods for testing the statistical similarity of studies. The tests assume that statistics for the studies are given as a normal or log normal distribution.

2.6.1 Tests of means

To test the similarity of the means of K studies (or log means for log normal studies) an ANOVA (Analysis of variance) can be performed. The test is defined as (null and alternative hypotheses):

$$H_0: \mu_1 = \mu_2 \dots = \mu_K$$

$$H_a: \mu_i \neq \mu_j, \text{ for at least one pair } (i, j)$$

The test indicates significant different means when at least one mean is different enough with respect to the other means. ANOVA assumes that the within study variances are equal among all studies but the sample sizes can be unequal. The calculation of ANOVA is based on the Mean Sum of Squares between and within studies:

$$MS_{between} = \frac{1}{k-1} \sum_i n_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$$

$$MS_{within} = \frac{1}{N-k} \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2 = \frac{1}{N-k} \sum_j (n_j - 1) s_j^2$$

A test statistic is then created as

$$F = MS_{between} / MS_{within}$$

And the null hypothesis (equal means) is rejected at significance level α if $F > F_{crit} = F_{\alpha}(k-1, N-k)$ is the $\alpha \cdot 100$ percentile from the F distribution with k-1, N-k degrees of freedoms.

In Babar
Tests of means are described in section 5.2.1 and example 7.3.

2.6.2 Tests of variances

The Bartlett's test of equal variances (Snedecor, G. Cochran, W., 1989) is used to test the equality of variances of K studies (or variances of logarithmic data for log normal distributions). The test is defined as (null and alternative hypotheses):

$$H_0: \sigma_1^2 = \sigma_2^2 \dots = \sigma_K^2$$

$$H_a: \sigma_i^2 \neq \sigma_j^2, \text{ for at least one pair } (i, j)$$

And the null hypothesis H_0 is tested at a given significance level α .

In Babar

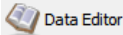
Tests of variances are described in section 5.2.1 and example 7.3.

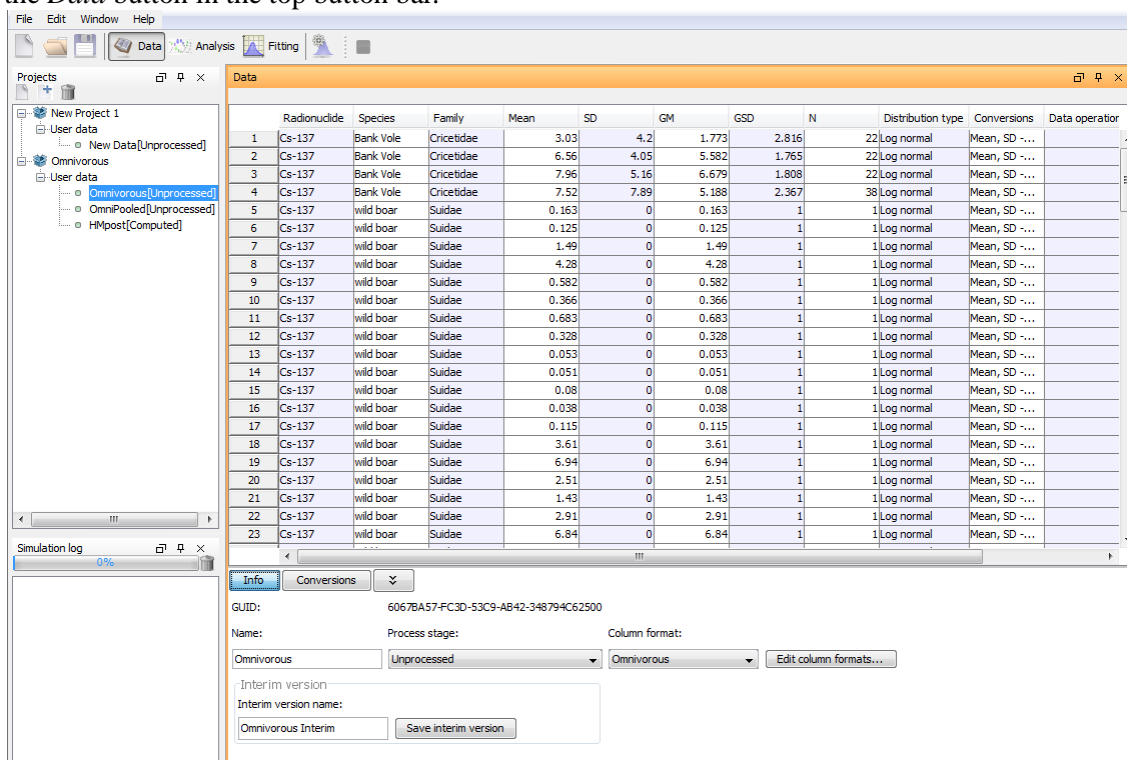
3 User interface – an overview

This document concerns the different parts of the Babar user interface. The user interface is based on different *views* which can be docked to the main window. All views can be set visible or hidden in the Window menu. The interface contains three *perspectives*: *The Data Editor perspective*, *The Analysis perspective* and *the Distribution Fitting perspective*. One can switch between the perspectives with the buttons in the toolbar.

Babar provide three default layouts, which provides the user with views and controls to edit data, perform analyses and perform distribution fitting. The layouts can be modified by dragging and docking the views as well as replacing views. The layouts can be reset to the default layouts from the Window->layout menu.

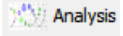
3.1 The Data Editor perspective

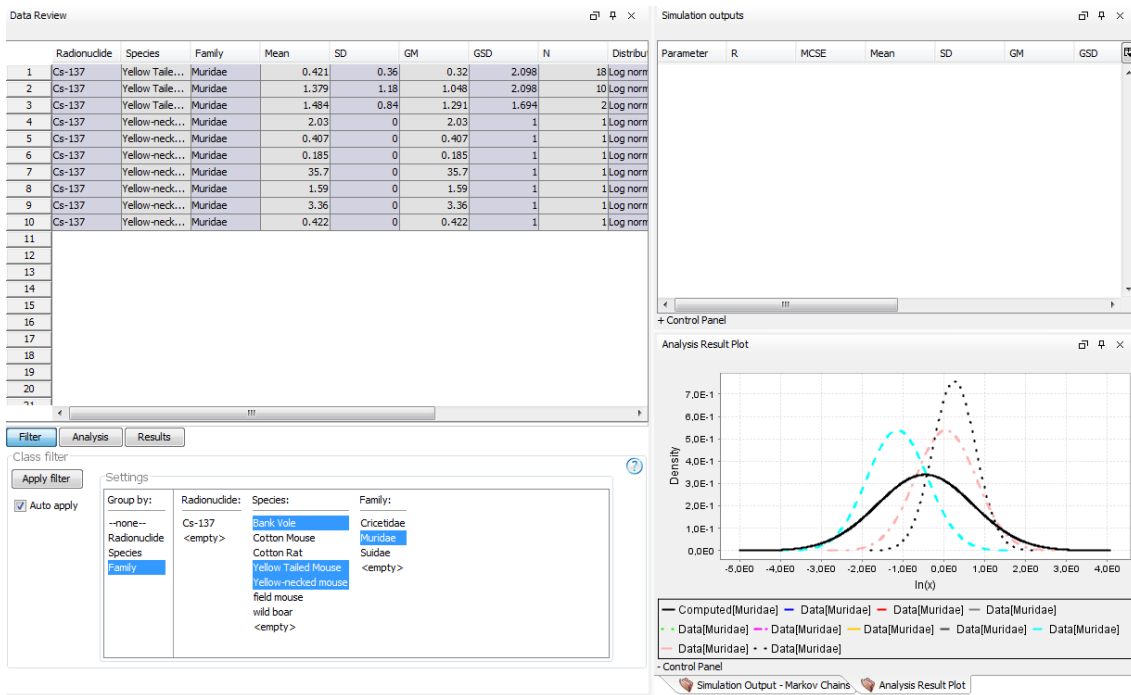
The data perspective (Figur 3-1) is available from the toolbar button  and contains a data editor view in which data can be edited. It also contains controls for editing information about the data sheet as well as performing conversions on statistics and units in the data. All (manual) changes to a data set are made in this view. The data perspective is opened by clicking the *Data* button in the top button bar.



Figur 3-1 The data perspective contains a data editor view and controls to display and modify information about the data as well as performing conversions on the data.


3.2 The Analysis perspective

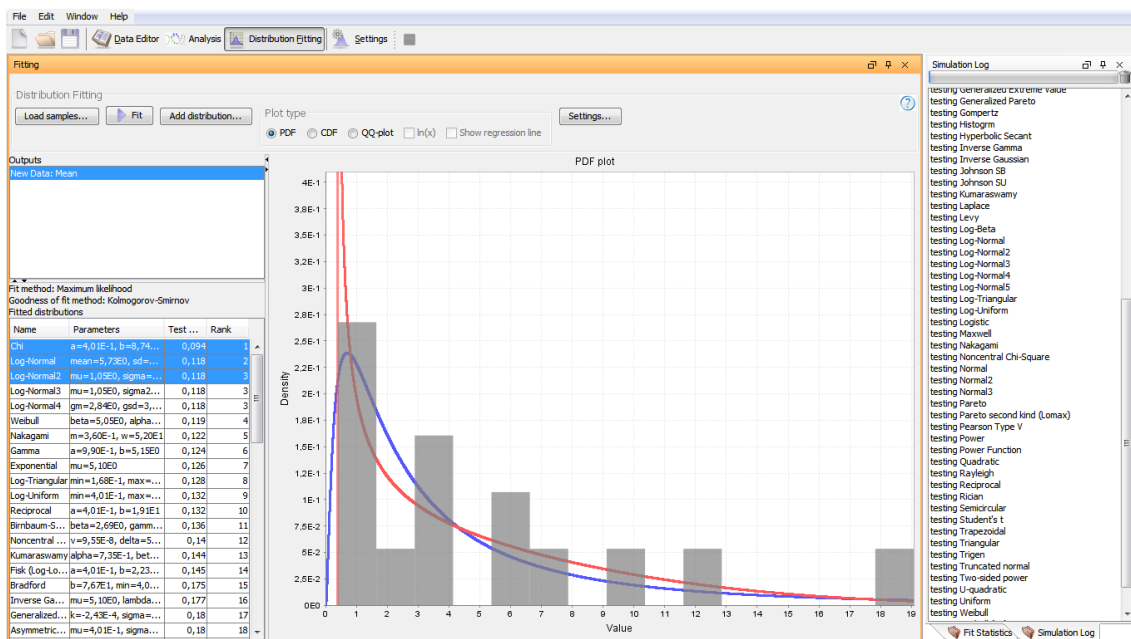
In the Analysis perspective (Figur 3-2) is available from the toolbar button  and is where all calculations are performed. Also, data and results can be inspected by showing graphs of all or a subset of the rows in the data. The sub view of the analysis perspective is described below.



Figur 3-2 The data analysis perspective with its sub views Data editor view, Simulation output view and Analysis results chart view.

3.3 The Distribution Fitting perspective









The Distribution Fitting perspective (Figur 3-3, available from the toolbar button ) contains views to fit probability distributions to measurements and inspect summaries of measurements and fitted distributions (such as mean, variance and percentiles of measurements).



Figur 3-3. The distribution fitting perspective.








3.4 The toolbar

The main toolbar provide access to shortcuts to the following functions:

	Create a new project
	Open an existing project
	Save the current project
 Data Editor	Open the Data Editor perspective
 Analysis	Open the Analysis perspective
 Distribution Fitting	Open the Distribution Fitting perspective
 Settings	Open the Settings dialog window
	Stop the current computation/simulation

3.5 The menus

3.5.1 The file menu

 New	Ctrl+N	Creates a new project
 Open	Ctrl+O	Opens an existing project
Recent Files		The ten most recently opened projects
 Close Project		Close the current project
 Save	Ctrl+S	Save the current project
 Save As	Ctrl+Skift+S	Save the current project as
Project Properties...		Opens the project properties (Name, author and description).
 Settings	F12	Opens the settings dialog window
 Exit		Exit Babar

3.5.2 The Edit menu

The edit menu has global entries for editing and removing the selected item (project or data).

3.5.3 The Window menu

The Window menu has entries for opening any of the views or switching between the three predefined perspectives/layouts (Data Editor, Analysis and Distribution Fitting). There are also functions for resetting the three perspectives to their predefined layouts.

3.5.4 The Help menu

Through the help menu, the user guide can be accessed, updates can be automatically downloaded and example data sets can be opened.

4 Creating and managing data sheets

All operations and calculations in Babar require data to be entered in a data sheet which is attached to the Babar project. In order for Babar to interpret the data correctly, the data must be entered in a well-defined way. A data sheet therefore has an associated *column format* that defines which data type are possible for each column.

4.1 The project view

The project view (Figure 4-1) displays all opened projects and data belonging to the projects. All other views reflect the project and /or data which is currently selected in the project view. From this view, data can be added to a project. The view provides a toolbar with shortcuts to create a new project, add new data to the selected project and delete the selected item. A context menu can be brought up by right clicking an item. From here, the project properties can be opened (allowing editing of the name of the project as well as the name of the author and comments for the project).

A project can be exported to excel. When exporting a project to excel, all data sheets are exported to a corresponding sheet in the excel file. The first excel sheet contains the properties of the project (name, author and comments).

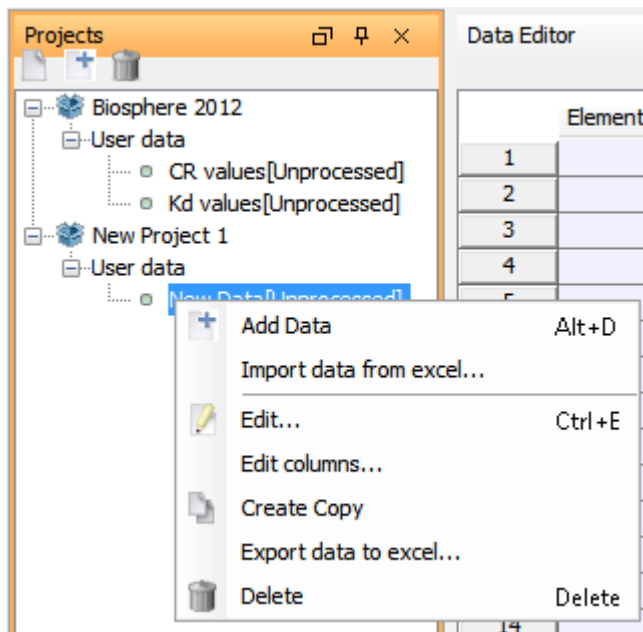
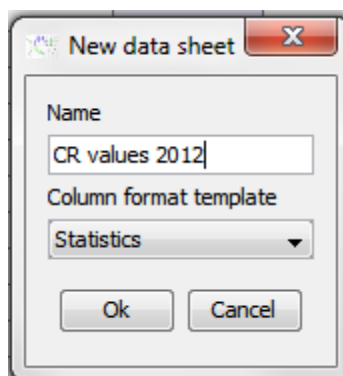


Figure 4-1. The project view.

4.2 Adding data to the project

A new data sheet is created by clicking the **+** icon in the projects view or right clicking the project and selecting Add Data. When creating a new data sheet, Babar asks for a name and a column format to use for the data sheet (See Figur 4-2. The dialog window asking for name and column format of the new data sheet Figur 4-2). The column format can be one of three build in formats or assigned a column format from an existing data sheet. The build in column formats are: Statistics and Raw data. The Statistics column format contains columns that are commonly used for representing multiple studies with statistics and an associated measurement distribution

type. The Raw data format is used for representing raw measurement values (e.g. for fitting distributions). After creating a data sheet, the column names and types can be modified in the column format editor (section 4.6.1). The column format editor is most easily opened by right clicking the data sheet in the project view and selecting “Edit columns...”.



Figur 4-2. The dialog window asking for name and column format of the new data sheet

Data in a data sheet is edited by selecting the sheet (by left clicking it in the projects view or right clicking and selecting Edit). The data sheet is then shown in the data editor view (section 3.1).

4.3 Importing a data sheet from excel

A data sheet can be imported from excel from the menu File->import or from the context menu in the project view. The excel file can contain several excel sheets and the imported names will be the same when imported in Babar.

Babar tries to interpret the content of the excel sheets as follows: The first row of each column is interpreted as the title of the column. If the name of a column matches any of the reserved column names used by Babar, the data type of the imported column will be guessed by Babar to be that corresponding to the reserved name.

The reserved column names are: Mean, N, SD, GM, GSD, Min, Max, Estimate, Info, Reference, Unit, Nominal, Value, Distribution type, Distribution, Conversions, Data operations, Detailed data operations.

If the title is not any of the reserved names but is non-empty and the content of the second cell in that column (i.e. the first value in the column) can be interpreted as number, the data type of that column is set to Observed Value. If the value of the second cell is not a number, the data type is set to Classification. Value of type classification is textual and can be used to filter studies to use in computations.

The automatic interpretation of data types performed by Babar can be changed after importation in the Settings->Column Settings.

4.4 Export data to excel

A data sheet can be exported to excel. This is done by selecting the data sheet in the projects view and select Export data to excel in the context menu.

4.5 Export project to excel

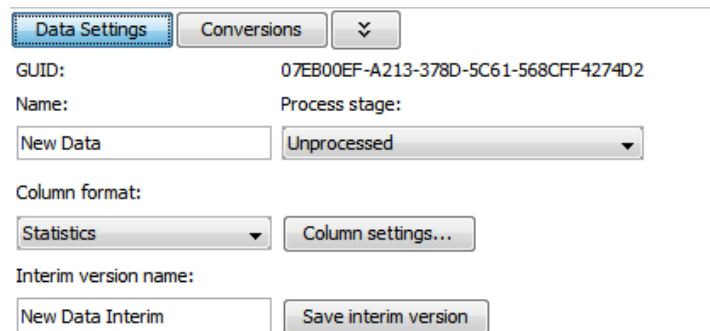
All data sheets in the project is exported to excel by selecting the project and select Export project to excel in the context menu. Each data sheet in the project is given a sheet in the excel file. The first sheet of the excel file will contain the properties of the project (name, author and comments).

4.6 Editing a data sheet

Data sheets are edited in the Data Editor View (by default visible in the Data Editor Perspective, section 3.1). A data sheet is shown as columns and rows and can be edited by entering values directly into the cells or pasted from excel or tab separated text files. Data is copied/pasted into data sheet by copying data from an excel or text. In Babar, go to the cell where the content is to be pasted and select Paste from the context menu (made visible by right clicking the cell in Babar) or press Ctrl-V. Content can be copied from the data sheet by selecting the content and selecting Copy from the context menu or pressing Ctrl-C.

The type of content of each column in a data sheet is restricted to the column's *data type*. The data types are defined in the Column Format editor, available from the context menu or the Data Settings panel (available from the Data Settings tab below the data editor).

The control panel (Figur 4-3) of the data editor contains controls for changing properties of the selected data sheet. The properties that can be changed here are: The name of the data sheet, the process stage of the data sheet, The column format (e.g. name, order and data types of the columns) and controls for performing conversions of statistics and units.



The screenshot shows the 'Data Settings' control panel. It features a 'Data Settings' tab, a 'Conversions' tab, and a dropdown arrow. Below these are fields for 'GUID' (07EB00EF-A213-378D-5C61-568CFF4274D2) and 'Name' (New Data). A 'Process stage' dropdown is set to 'Unprocessed'. Under 'Column format', there is a dropdown set to 'Statistics' and a 'Column settings...' button. At the bottom, there is an 'Interim version name' field (New Data Interim) and a 'Save interim version' button.

Figur 4-3. The control panel of the data editor.

4.6.1 Column Settings

The column settings is opened by selecting “Edit columns...” context menu in the project view or in the data editor view (opened by right clicking somewhere in the data sheet) or from the Settings window.

Modifying the name, data type and order of individual columns

Select the name of the data sheet in the drop down list. To add a new column to the selected column format, click “Add column”. To modify the data type a column, select the row corresponding to the column and double click the cell in the first column (“Column type”). To change the name of a column, select the row corresponding to the column and double click the cell in the second column (“Column name”) and change the name. The name of the column must be unique among all columns in the columns format.

The column type decides which data the column can hold and how the data is interpreted by Babar. For instance, a column of the type Men can only hold numbers and will be interpreted by

Babar as the arithmetic mean value. The following column types are available (columns marked with asterisk can only occur once per column format/data sheet):

Statistics

- Mean (The arithmetic mean value)*
- SD (The arithmetic standard deviation)*
- GM (The geometric mean)*
- GSD (The geometric standard deviation)*
- N (Sample size or weight)*
- Min (observed minimum value)*
- Max (observed maximum value)*
- Nominal (A nominal value or best estimate)*
- Value (Some value, e.g. for use for values of regression variables)

Classification, distributions, Units and references

- Classification (Textual value, used to classify data. Needed in most calculations)
- Distribution type (Normal or Log normal, necessary for all the computations to interpret the statistics correctly)-*
- Distribution (A distribution with specified parameters, used for the weighted resampling method)*
- Unit (The unit, e.g. Bq/Kg.)*
- Reference (Textual value representing references for the parameter)

Columns with information written by Babar after computations. It is highly recommended that these are included for data used by computations.

- Conversion info (Computed by Babar, holds information about performed conversions)*
- Data operation info (Computed by Babar, holds brief information about which operations led to the data in the row)*
- Detailed data operation info (Computed by Babar, holds detailed information about which operations led to the data in the row)*

Column formats

Formats:
 New Data

Add column Move up Move down Delete column

Column type	Column name
Classification	Element
Classification	Species
Classification	Classification
N	N
Mean	Mean
SD	SD
GM	GM
GSD	GSD
Distribution type	Measurement distribution
Reference	Reference
Unit	Unit
Distribution	Distribution
Conversions	Performed conversions (computed)
Data operations	Performed operations (computed)
Detailed data operations	Detailed data operations (computed)

Figur 4-4. The column format settings panel

4.6.2 Conversions

Babar provides functionality to convert some statistics to other statistics. Conversion between statistics is necessary when Babar requires data to conform to certain format. For instance, most calculations requires measurements with a log normal measurement distribution to be given as geometric means (GM) and geometric standard deviations (GSD). Performed conversions are logged in the column “Performed Conversions”. **Note: The column format of the data sheet must contain a column of the type “Performed Conversions “ in order to record any performed conversions.**

Conversions are performed from the Conversion Panel, available from the Conversion Tab in the Data editor view. To perform a conversion, select the missing statistics from the list of “Missing Statistics” and select a conversion path.

The conversion paths available in Babar are shown in Table 4-1. When a conversion is performed, the statistic in the column “Missing Statistic” is calculated from the statistics in the column “Available statistics” using the formula in the “Formula” column. The conversion is possible only under the measurement distribution in the “Distribution” column. **Note: The conversions assume that the measurements conform perfectly to the specified measurement distribution and are in general approximate. For example, the resulting GM,GSD (converted from Mean,SD) do not equal the GM,GSD calculated from the original measurements. Instead they are the GM,GSD of the log normal distribution with given Mean and SD.**

Table 4-1. Available paths for converting between statistics.

Missing statistic	Available statistics	Formula	Distribution
GM	Mean,SD	$GM = \frac{mean}{\sqrt{\left(\frac{sd}{mean}\right)^2 + 1}}$	log normal

GSD	Mean,SD	$SD = \exp\left(\sqrt{\ln\left(\left(\frac{sd}{mean}\right)^2 + 1\right)}\right)$	log normal
Mean	GM,GSD	Mean = $\exp(\mu + 0.5\sigma^2)$	log normal
SD	GM,GSD	SD = $\exp(\mu + 0.5\sigma^2) \times \exp\sigma^2 - 1$	log normal
Nominal	GM	Nominal=GM	Normal, Log normal
Nominal	Mean	Nominal=Mean	Normal, Log normal
GM	Min,Max	$GM = max \times min$	Log normal
GSD	Min,Max	$GSD = \left(\frac{max}{min}\right)^{1/(2 \times 1.96)}$	Log normal
Mean	Min,Max	$Mean = \frac{max + min}{2}$	Normal
Mean	Min,Max	GM, GSD from Min,max above Mean from GM,GSD above	Log normal
SD	Min,Max	$SD = \frac{max - min}{1.96}$	Normal
SD	Min,Max	GM, GSD from Min,max above SD from GM,GSD above	Log normal

4.6.3 Units and unit conversions

Babar can store units of studies. Units can be entered in a column which is of the data type “Unit”. Conversion can also be performed between different units (e.g. Kg to g or Bq/g to Bq/Kg) by user defined conversion rules.

Examples of correctly formatted unit strings: Kg, Bq/Kg, (Bq/ Kg), (Bq /KgDw)/ (Bq/KgFw)
Example of incorrectly formatted units C* Bq/Kg, 0.2*Bq/Kg, Bq/Kg).

Babar supports a simple interface for conversions of units (e.g. Bq/Kg, Bq/g) by user-defined rules. Performed conversions are logged in the column “Performed Conversions”. **Note: if the column “Performed Conversions “ does not exist in the column format, the conversions are not logged in any way.**

Units and conversion rule are defined in the Unit settings editor (Figur 4-5, available from the settings dialog window). Units and conversion rules are stored in the project.

Units

Available units

- Bq/Kg
- Bq/g

Edit unit (e.g. Bq/Kg)

Bq / Kg

Add Delete

Add all units used in project

Unit conversions

Available conversions

- Conversion: Bq/Kg * 0.
- Conversion 1: Bq/g * 10

Edit conversion (e.g. Bq/g * 1000 = Bq/kg)

Bq/g * 1000.0 = Bq/Kg

Name: Conversion1

Add Delete

Figur 4-5. Unit and unit conversion settings available from the settings dialog window.

4.6.4 Process stages

A data sheet can be assigned a process stage to prohibit editing of the data. Each process stage prohibit or allows different operations on the data, for example manual editing, conversions or modifications in the data sheet by computations. The process stage “Unprocessed” is the default stage and pose no restrictions on the data. The use of process stages is optional, but all computations in Babar automatically sets the result data to Computed, to disallow manual editing of the data sheet. The available process stages are show in Table 4-2.

Table 4-2. Process stages

Stage	Allowed modifications	Use
Unprocessed	All.	For new data sheets
Preprocessed	Conversions, computation info can be added by Babar.	Locked for manual editing
Processed	Computation info can be added by Babar.	Locked for conversions. To be used in computations
Computed	None.	Computed data. Results from computations have this stage as default.
Postprocessed	Conversions.	Computed data for preprocessing (allowing conversions of statistics and units).

5 Performing computations

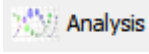
Computations are performed in the Analysis perspective . The default layout of this view provide the following views:

Table 5-1. The views in the Analysis perspective.

View	Description
Analysis view	Shows the selected datasheet as write protected, for review prior to computations. A filter based on values of classification columns allows subsets of studies to be included in computations (e.g. specific species or sources of studies). A control panel provide controls for performing computations on the filtered studies. Results from computations summarized as statistics can be reviewed and exported to a new or existing data sheet.
Analysis Result Chart	A chart showing the selected studies as PDFs.
Simulation outputs	A table showing detailed summary statistics of and convergence quantities of the parameters from Bayesian computations.
Simulations – MCMC Chart and Bar chart.	Graphs based on the raw simulation samples for Bayesian simulations for the simulation outputs selected in the simulation output table. The samples can be viewed as a Markov Chain chart for inspecting convergence of a simulation or bar charts.

5.1 The Analysis data view

The Analysis view (Figur 5-1) provide functionality for reviewing subsets of studies/rows and performing computations, tests and analysis on selected studies/rows. Data from the current data sheet is shown in a table but cannot be manually modified. The bottom panel provides controls for filtering out studies (i.e. rows of a data sheet) and controls for performing operations on the data.

5.1.1 Switching between the current data or result of computations

The table is used both to show the data used for computations or the results of the current simulation. Two buttons control if the table shows the filtered data used in a computation or the results of the current computation:

Selected data sheet: Shows the rows of the selected data sheet that have been filtered and to be used in a computation.

Current simulation result: Shows the results of the latest simulation (pooling or Bayesian simulations). The results can be exported to a data sheet from the Result Tab.

The control panel in the bottom of the view has three tabs: Filter, Analysis and Results.

5.1.2 The Filter tab

The filter tab (available by selecting the “Filter” tab) provide controls for filtering out studies in the selected data sheet. The filter shows lists where values of each categorical column can be selected. Each selection updates the filter and the rows matching the filter are shown in the data table. The “Group by” list shows the names of the categorical columns of the data sheet. The selection of column name does not change the filtered columns, but defines how computations will interpret rows with the same values of the filtered columns.

Example: Filtering studies for combination or pooling of studies

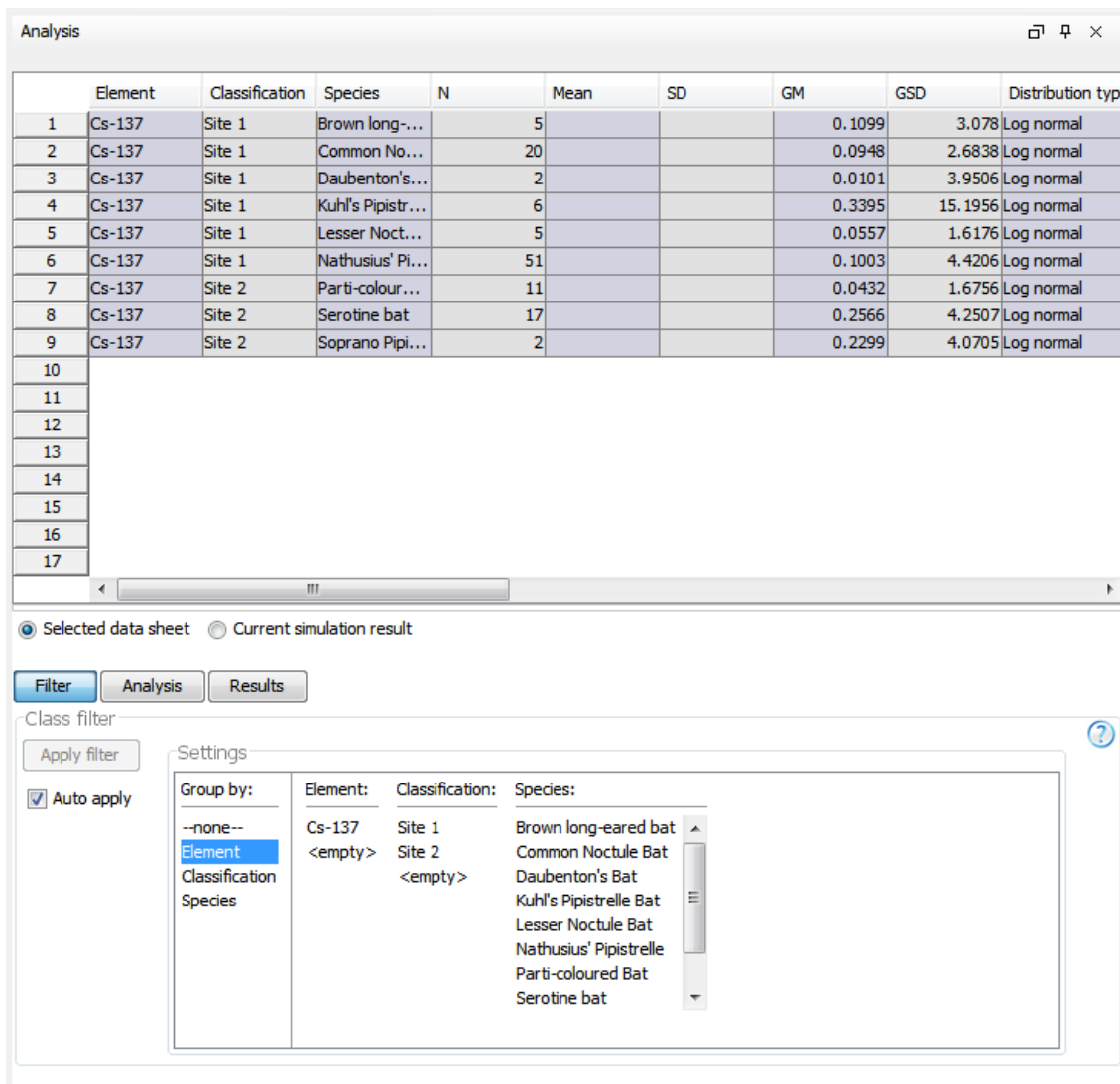
The data shown in figure 7 contains nine studies of Cs-137 in different species of Bats from the two sites. Assume now that the studies are to be combined or pooled per site, the “Group by” column is then selected as “Site” the computation algorithm can perform the combination as wanted. If the value of the “Group by” column is instead selected as “Element” then all nine rows are combined, irrespective of the value of the site.

Example: Filtering studies for hierarchical updating

The value of the “Group by” column has a slightly different interpretation for hierarchical updating than for combinations or pooling. The “Group by” value is then used to define which studies are to be treated as separate groups. Selecting “Species” as the value of the “Group by” column would estimate all nine species hierarchically. If only species from site 1 would be included, then a filter must be set to include only those rows.

Example: Filtering studies for testing mean and variances

If the mean values of the studies from the same site were to be tested, then value of the “Group by” column should be set to “Site” (similarly to for combinations and pooling of studies).



Figur 5-1. The Analysis view showing the data table and the class filter.

5.2 The Analysis tab

The analysis tab provide controls for performing computations and tests of the filtered rows.

5.2.1 Test mean/variances

The “Test mean/variances” tab provide controls for performing statistical tests of multiple means and/or variances. If the studies have a log normal measurement distributions, the test is performed using the geometric means and variances (i.e. $\ln(\text{GM})$ and $\ln(\text{GSD})$).

Test of equal means: Test mean values with ANOVA.

Test of equal variances: Test variances with Bartlett’s test.

Exclude rows with missing statistics: If selected, excludes rows that does not sufficient statistics.

Alpha: Specifies a significance level. The value only affects the message at the end of the test, but not the test itself or reported p-values.

Test: Performs the selected tests of mean and/or variances. The results are displayed in the simulation log window.

5.2.2 Pooling

The “Pooling” tab provides controls for performing pooling of means and/or variances.

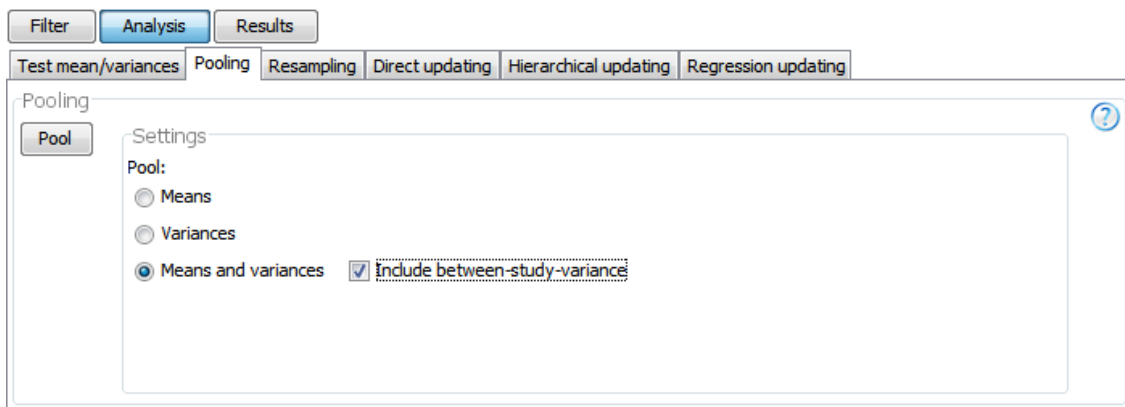
Pool means: Pool means of the selected studies.

Pool variances: Pool variances of the selected studies.

Pool means and variances: Pool both mean and variances.

Include between-study-variance: Calculates the combined mean and variances which includes the estimate of the between study variance.

Pool: Computes the pooled/combined mean and/or variances and. The resulting statistics are shown in the current simulation result table.



5.2.3 Resampling

The resampling panel has two buttons:

Generate samples: Generates samples from the probability distributions defined for the selected rows. After successful simulation, the generated samples will be visible in the Distribution Fitting tool where probability distributions can be fitted to the samples and statistics can be calculated for the samples.

Simulation settings...: Opens the simulation settings window. Here, the number of samples to obtain from the probability distributions can be set.

5.2.4 Direct updating

The *Direct updating* panel provides controls for performing Bayesian updating using the conjugate prior or semi-conjugate prior methods.

Conjugate prior: Interpret the prior as a joint-conjugate prior.

Semi-conjugate prior: Interpret the prior as a semi-conjugate prior.

Class column: Which column contains values to distinguish between prior and observed data rows.

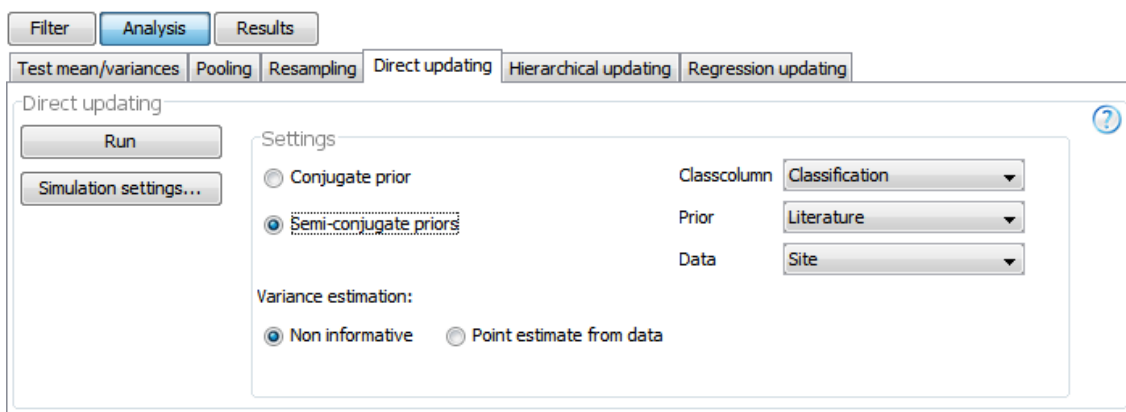
Prior: The value of the class column that specifies rows with prior data (e.g. “Literature”).

Data: The value of the class column that specifies rows with observed data (e.g. “Site”).

Non informative prior (only for semi-conjugate prior): Estimate the posterior variance using a approximate non-informative prior. The variance is then considered uncertain and mainly estimated from observed data but adjusted to account for the updated posterior mean.

Point estimate from data (only for semi-conjugate prior): Considers the variance as known and equal to a point estimate of the sample variance from observed data.

Run: Performs the computation of the posterior distributions. The resulting posterior distributions and convergence statistics are summarized in the Simulation output view. Information about simulation parameters and convergence statistics is shown in the Simulation Information view. Statistics of the predicted distribution which uses point estimates of the mean and variance (or GM,GSD) are shown in the current simulation result table.



5.2.5 Hierarchical updating

The hierarchical updating panel contains controls to perform hierarchical updating of the mean values of the filtered studies. There are two options for how to estimate the variance parameter:

Point estimate from data: The variance is considered known and equal to the variance of each study.

Common variance with non-informative prior: All groups are modeled as having the same true variance. It is considered uncertain and estimated using a non-informative prior.

Simulation settings...: Opens the simulation settings dialog window. Here, the number of samples can be selected as well as parameters specific for the Markov Chain Monte Carlo simulations.

Run: Starts the posterior simulation. Performs the computation of the posterior distributions. The resulting posterior distributions and convergence statistics are summarized in the Simulation output view. Information about simulation parameters and convergence statistics is

shown in the Simulation Information view. Statistics of the predicted distribution which uses point estimates of the mean and variance (or GM,GSD) are shown in the current simulation result table.

5.2.6 Regression updating

The *Regression updating* panel contains controls to perform updating of the parameters (coefficients) of a linear regression model. The parameters are considered uncertain and assigned prior distributions. After updating, the posterior distributions of the parameters account for both the prior distribution and the observed data of a dependent (response) variable and the independent variables.

There are two tables:

Observed variables: The dependent and independent variables are mapped to columns of measurements in the selected data sheet.

Prior distributions: Prior distributions are defined for each of the parameters (coefficients) of the regression model. As default, approximate non-informative prior distributions are defined, centered at 0 and with standard deviation $1e6$.

Run: Performs the computation of the posterior distributions. The resulting posterior distributions and convergence statistics are summarized in the Simulation output view. Information about simulation parameters and convergence statistics is shown in the Simulation Information view.

5.3 Reviewing results from computations

5.3.1 The Analysis view

After successful computations of any of the Pooling or Bayesian methods, the resulting statistics are exported to a temporary data sheet which is shown in the Analysis view (Figur 5-2). This sheet is visible per default in the Analysis data table view directly after any computation has finished. The two buttons under the analysis table can be used to switch between showing data from the selected data sheet and the results from the latest (current) computation.

The resulting data sheet stores information about the performed simulation. When a row is selected in the Analysis view, the data used for calculating that row is shown in the Data Information view.

	Element	Classification	Species	GM	GSD	N	Performed operations (comput
1	Cs-137	Site	Brown long-eared bat	0.1007	3.078		Hierarchical updating
2	Cs-137	Site	Common Noctule Bat	0.0936	2.6838		Hierarchical updating
3	Cs-137	Site	Daubenton's Bat	0.0417	3.9506		Hierarchical updating
4	Cs-137	Site	Kuhl's Pipistrelle Bat	0.1287	15.1956		Hierarchical updating
5	Cs-137	Site	Lesser Noctule Bat	0.0584	1.6176		Hierarchical updating
6	Cs-137	Site	Nathusius' Pipistrelle	0.0989	4.4206		Hierarchical updating
7	Cs-137	Site	Parti-coloured Bat	0.0451	1.6756		Hierarchical updating
8	Cs-137	Site	Serotine bat	0.2021	4.2507		Hierarchical updating
9	Cs-137	Site	Soprano Pipistrelle Bat	0.1212	4.0705		Hierarchical updating

Shown data

Selected data sheet
 Current simulation result

Figur 5-2. The table showing the current simulation result in the Analysis view.

5.3.2 The Analysis Result Chart View

The summary statistics from a data sheet or result sheet can be plotted in the Analysis Result Chart view (Figur 5-3). Here, the predicted probability distributions (normal distributions fitted to the point estimates of the posterior distributions of μ and σ) are plotted for selected studies or rows of the result sheet. For rows with normal measurement distributions, the point estimates of μ and σ are taken from the Mean and SD columns of the result data sheet. For rows with log normal measurement distributions, the point estimates are taken as $\ln GM$ and $\ln GSD$.

Note: The Analysis Result Chart only plots data from rows of the Analysis view and not from the Data Editor. If the Analysis Result Chart shows data from the Current Simulation, then the Analysis Result Chart displays rows from the Current Simulation.

The control panel of the view has the following controls:

Legends: Legends can be turned on and off from the control panel of the view and the columns to include in the legend can be selected from the list. If the full IDs button is selected, all columns are used to construct the legends.

Shown data - Data used in computed rows: If checked and the selected row has content in the Detailed Simulation Info column, then the data used to compute this row is also plotted.

Shown data – Selected rows only: If checked, shows the rows selected in the table in the Analysis view. If unchecked, all rows visible in the Analysis view is plotted.

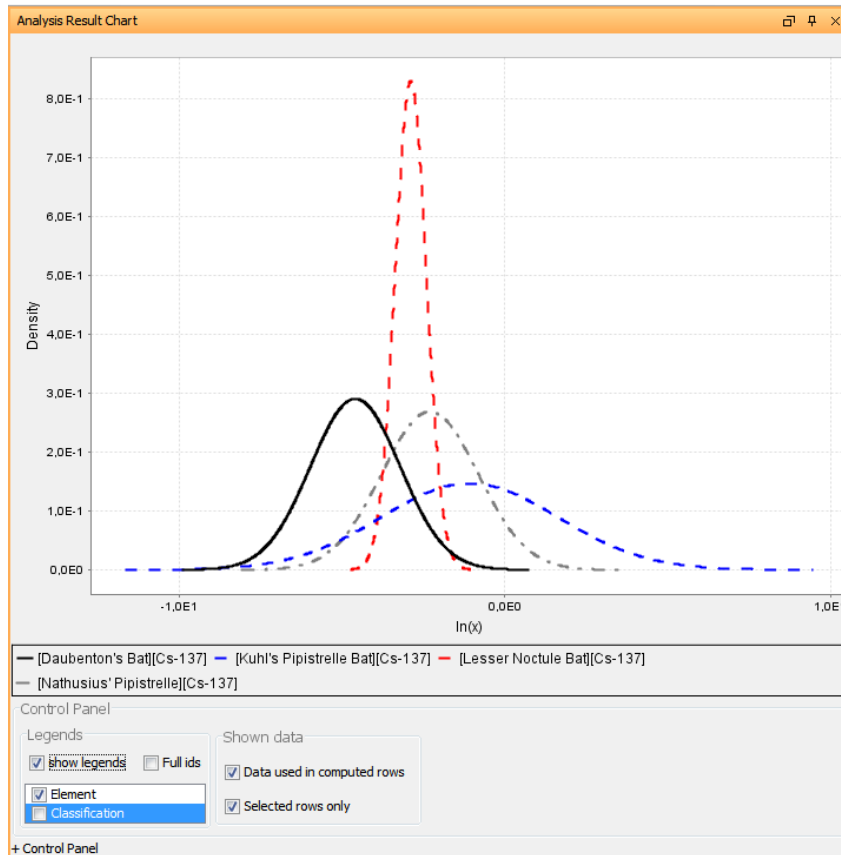


Figure 5-3. The Analysis Result Chart view.

5.4 Inspecting results and convergence diagnostics from Bayesian simulations.

Bayesian computations are based on simulations using a finite number of simulated draws from the posterior distributions of the model parameters. The simulation outputs should therefore be inspected for convergence before any estimates is used. The following views can be used to inspect the results from Bayesian simulations: The Simulation Output Statistics, the Simulation Output Chart View and Simulation Information View.

5.4.1 Simulation Output Statistics table view

The Simulation Output Statistics table shows summary statistics of all simulation outputs.

The following parameters are reported for a Bayesian simulation:

Direct updating:

Mu: The mean (μ) of the model.

Sigma^2: The variance (σ^2) of the model.

Hierarchical updating:

Mu[j] : The mean (μ_j) of the model for group j.

Sigma^2[j]: The variance (σ_j^2) of the model for group j.

Mu.pop: The mean (μ_{pop}) of the population/prior distribution of a hierarchical model.

Sigma.pop^2: The variance (σ_{pop}^2) of the population/prior distribution of a hierarchical model.

Mu.pred: The predictive population mean (μ_{pred}) of a hierarchical model, simulated as $N(\mu^*, \sigma^{2*})$ where μ^*, σ^{2*} are samples from the posterior distribution.

Regression updating (with the model $y = intercept + b_1 \cdot X_1 + \dots + b_k + X_k$):

Intercept: The intercept parameter

b_k : The k:th parameter (coefficient) of the regression model

Sigma^2 (σ^2): The squared model error.

Note: For log normal measurement models, the simulation outputs are generally on a log-transformed scale. That is, the parameter mu (μ) and sigma^2 (σ^2) denotes the mean and variance of log-transformed measurement variable.

The table has the following columns:

N: The total number of simulated samples.

R: The Gelman Rubin convergence statistic. Convergence is said to be reached if R is close to 1. One interpretation of R is that it is the potential reduction of the scale (“width”) of the posterior distribution that is possible if more samples are collected.

MCSE: The Monte Carlo Standard Error of the Mean. The MSCE quantifies the precision of the mean of the posterior distribution due to the limited number of samples. It can be used to give the number of the decimals that can be reported (of the mean).

Mean: The value of the posterior distribution.

SD: The standard deviation of the posterior distribution

GM: The geometric mean of the posterior distribution

GSD: The geometric standard deviation of the posterior distribution

Percentiles: Percentiles the posterior distribution.

Output	N	R	MCSE	Mean	SD	2.5% percentile	5% percentile	50% percentile	95%
mu[Brown long-eared bat]	28 500	1.00E0	2.83E-3	-2.24E0	4.77E-1	-3.19E0	-3.03E0	-2.24E0	-1.45E0
mu[Common Noctule Bat]	28 500	1.00E0	1.69E-3	-2.34E0	2.86E-1	-2.91E0	-2.82E0	-2.34E0	-1.87E0
mu[Daubenton's Bat]	28 500	1.00E0	4.74E-3	-3.13E0	8.01E-1	-4.93E0	-4.60E0	-3.04E0	-2.01E0
mu[Kuhl's Pipistrelle Bat]	28 500	1.00E0	3.10E-3	-1.58E0	5.23E-1	-2.50E0	-2.38E0	-1.60E0	-6.70E-1
mu[Lesser Noctule Bat]	28 500	1.00E0	2.99E-3	-2.61E0	5.05E-1	-3.68E0	-3.48E0	-2.58E0	-1.84E0
mu[Nathusius' Pipistrelle]	28 500	1.00E0	1.12E-3	-2.30E0	1.89E-1	-2.66E0	-2.60E0	-2.30E0	-1.99E0
mu[Parti-coloured Bat]	28 500	1.00E0	2.45E-3	-2.87E0	4.14E-1	-3.70E0	-3.57E0	-2.87E0	-2.21E0
mu[Serotine bat]	28 500	1.00E0	2.09E-3	-1.58E0	3.53E-1	-2.27E0	-2.17E0	-1.57E0	-1.00E0
mu[Soprano Pipistrelle Bat]	28 500	1.00E0	3.80E-3	-1.98E0	6.42E-1	-3.16E0	-2.96E0	-2.03E0	-8.54E-1
mu.pop	28 500	1.00E0	2.09E-3	-2.29E0	3.53E-1	-3.03E0	-2.87E0	-2.28E0	-1.74E0
mu.pred	28 500	1.00E0	5.87E-3	-2.28E0	9.91E-1	-4.35E0	-3.88E0	-2.27E0	-7.35E-1
sigma.pop^2	28 500	1.00E0	6.05E-3	8.40E-1	1.02E0	1.97E-2	5.04E-2	5.49E-1	2.53E0
sigma^2	28 500	1.00E0	1.65E-3	2.00E0	2.79E-1	1.53E0	1.59E0	1.97E0	2.50E0

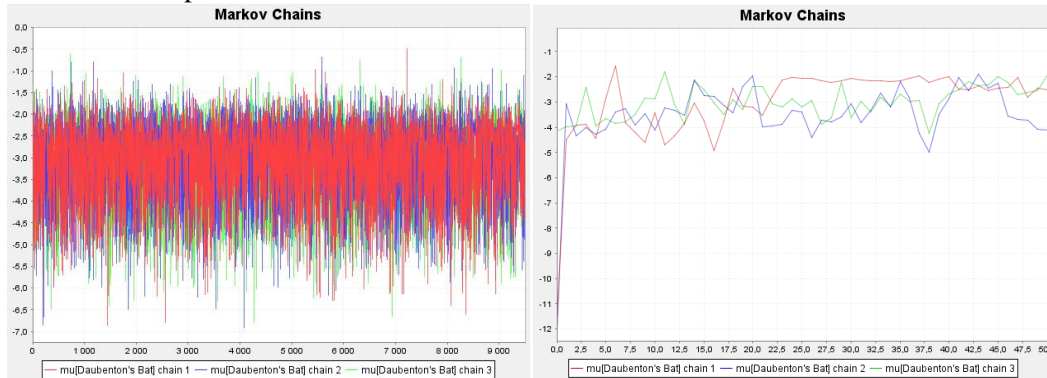
5.4.2 Simulation Output Chart View

The Simulation Output Chart view displays the posterior simulation samples for the output selected in the Simulation Output Statistics table view. The view has two chart types:

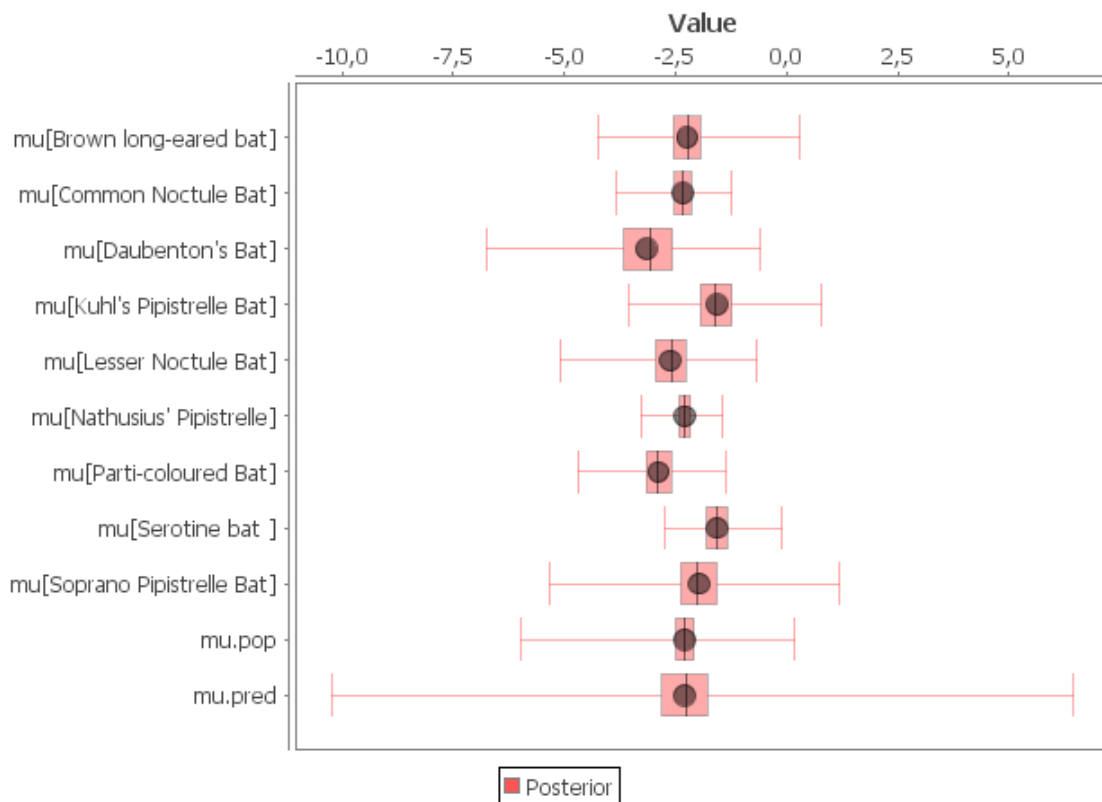
MCMC Chains: Displays the series of samples obtained from each independent chain of the Markov Chain Monte Carlo (MCMC) simulation. The chart is used to assess convergence of the simulations. Diverging chains of samples is a sign of insufficient convergence of the simulation. The impact of the random start values and the choice of burnin factor can also be assessed: If the first shown values of the chains are very different from the rest of the obtained samples, then a larger burn-in factor might be needed to exclude those values from the samples used in inferences. Figure 5-4 displays two MCMC charts. The first shows well mixed samples with no

patterns of diverging chains. The second chart shows an example of a simulation based on very few samples. There, the first few samples are seen to be affected by the random start values and a section of the iterations that are stuck in the posterior distribution.

Bar chart: Bars showing the 95% probability intervals (vertical line), 50% probability intervals (box), medians (black vertical line) and means (circles) are shown for the outputs selected in the Simulation Output statistics table.



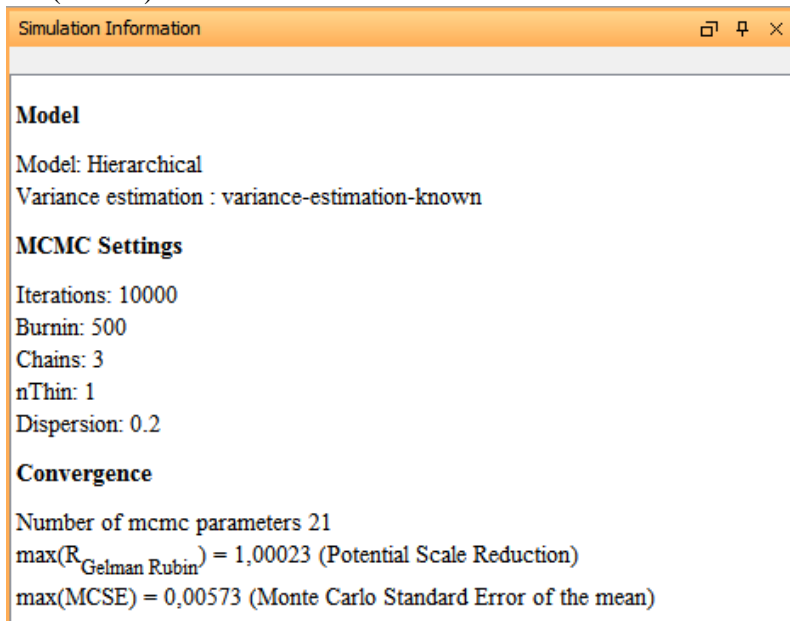
Figur 5-4. MCMC Chains Chart. *Left chart: The last 9500 samples of three chains excluding the first samples (burnin=500 samples). The chains are well mixed and show no sign of divergence or any large influence from the start values. Right chart: The 50 first samples from the same parameter when no samples are excluded (burnin=0) The first 2-3 samples are clearly seen to be affected by the random random start values. There is also a part (between iterations 23-40) of the sample of one of the chains which is stuck around the value -2 in the posterior distribution. To get samples that are more independent of the random start values a larger burnin factor should be choosen. To get samples that covers a larger part of the posterior distribution, a larger number of iterations should be run.*



Figur 5-5. The Simulations Bar chart view displaying summaries of posterior samples of selected parameters


5.4.3 Simulation Information View

The Simulation information view summarizes the settings and some convergence measures of the last (current) simulation.



Figur 5-6. The Simulation Information view summarizes the simulation settings and convergence measures from the latest simulation (here a hierarchical model).

6 The Settings window

The settings view are accessible from the toolbar button . The following is a description of the different pages of the Settings View.

6.1.1 Application Settings

Here, the automatic check for updates on startup can be turned on/off. Also, database login settings are set here.

Application Settings

Check for updates on startup

Database

Driver:

Url:

Login name:

Login password:

Figur 6-1 Application Settings. These settings are stored in the application (not in the project). Here the database server and connection strings can be set.

6.1.2 The project properties

In the Project properties, the name, author and description for the project are set.

6.1.3 Column Format Settings

Here, the column formats used for the project's data can be inspected and/or modified. New Data formats can be created or cloned from existing formats for use in new data sheets. Formats can be exported and imported to be shared among users/projects. **Note:** Modifying a column format will also modify the structure of all data sheets using that column format.

Column formats

Statistics ▾ Apply in current sheet

Statistics Add Column Format Remove Column Format

Export... Import...

Clone

Add column Remove

Move up Move down

Column type	Column name
Classification	Element
Classification	Species
Classification	Classification
N	N
Mean	Mean
SD	SD
GM	GM
GSD	GSD
Distribution type	Measurement distribution
Reference	Reference
Unit	Unit
Distribution	Distribution
Conversions	Performed conversions (computed)
Data operations	Performed operations (computed)
Detailed data operations	Detailed data operations (computed)

Figur 6-2 Column Format settings.

6.1.4 Unit Settings and Unit Conversion Settings

Here, units and unit conversion rules are defined. Units can be imported from the project's data sheets. It is only necessary to define units if these are to be converted by Babar. Rules for Unit conversions are defined as the simple mapping of one unit times a scalar number.

Units

Available units	Edit unit (e.g. Bq/Kg)
Bq/Kg Bq Kg (Bq/Kg)/Bq(Kgfw) Bq(Kgfw)	Bq/Kg / Bq(Kgfw) Add Delete Add all units used in project

Unit conversions

Available conversions	Edit conversion (e.g. Bq/Kg * 1.234 = Bq/m2)
Conversion: Bq/Kg * 0.	Bq/Kg * 0.2 = /Bq(Kgfw) Name: <input type="text"/> Add Delete

Figur 6-3 Unit settings and Unit conversion settings.

6.1.5 Fitting Settings

Figur 6-4 shows the Fit settings page. Here, the method used to fit distributions parameters to data and the method used for testing the fit of the fitted distributions can be selected. The fit methods are:

Maximum Likelihood: Uses maximum likelihood method for fitting.

Below Limit Of Detection: Uses a method based on regression fitting to estimate the parameters of a normal or log normal distribution when the observed values have value below detection limit. Values below detection limit is entered in a data sheet by prepending the value with “<” (less than). The value of the setting *Show regression in QQ plot* indicates whether the regression line fitted with this method should be shown in the QQ plot.

The available Goodness of fit methods are:

Kolmogorov Smirnov: Uses the Kolmogorov Smirnov method.

Anderson Darling: Uses the Anderson Darling method.

The method to calculate the bin size of the samples histogram can also be changed here, or a custom bin size can be set.

Fit settings

Fit method

Maximum likelihood Below Limit Of Detection

Goodness of fit test

Kolmogorov Smirnov Andersson Darling

Distributions

Continuous Discrete

Alpha
Anglit
Arc-sine
Asymmetric Laplace distribution
Asymmetric double exponential
Beta
Beta (Generalized)
Beta-PERT

Add all
Clear all

Plot

Method Scotts method

Custom bin size

Show regression in QQ plot

Figur 6-4 The distribution fitting settings.

6.1.6 Simulation Settings

The Simulation Settings shows settings used for the Resampling method and Bayesian computations. For the Resampling method, only the *iteration* setting is used. The MCMC (Markov Chain Monte Carlo) settings are used for Bayesian methods.

Iterations: The number of iterations to run for each independent chain of samples for Bayesian simulations. The minimum number of samples to obtain for resampling simulation.

Burn in: The number of initial samples of each chain to use as burn-in in Bayesian simulations. The first burn in samples will not be recorded or used in any tables, statistics or charts. The purpose is to diminish the impact of the random start values of the simulations. A general conservative recommendation of some authors is to use half the iterations as burn-in. For the methods implemented in Babar (i.e. based on a semi-analytical Gibbs sampler) a smaller number.

Thinning factor: If set to K, only the Kth value will be recorded by the Bayesian simulations. The purpose is to avoid autocorrelation in the simulated chains. K=1 is the default and is sufficient for the methods implemented in Babar (i.e. methods based on the semi-analytical Gibbs Sampler)

Number of Chains: The number of independent chains to simulate for Bayesian simulations. The independent chains are used to assess the convergence of the simulated samples (by calculation of the Gelman Rubin convergence statistic R). The default value is three.

Dispersion: a factor determining the dispersion/variation of the random start values of the chains in Bayesian simulations. The default value is 0.1.

Estimation method: How point estimates are estimated from the posterior distributions. If set to “posterior median”, the medians of μ and σ^2 is used to estimate Mean, SD or GM, GSD of the predicted distributions. If set to “predicted distribution statistics” the predicted distribution is simulated using all obtained posterior samples for μ and σ^2 and point estimates of the mean/GM is taken as the mean/GM of the predicted distribution. The point estimate of SD/GSD is taken as the SD/GSD of the predicted distributions. The second method in general produces a larger SD/GSD than the first method.

Simulation Settings

Iterations	<input type="text" value="10000"/>
Markov Chains Settings	
Burn in	<input type="text" value="500"/>
Thinning factor	<input type="text" value="1"/>
Number of chains	<input type="text" value="3"/>
Dispersion	<input type="text" value="0.2"/>
Estimation method	<input type="text" value="Posterior median"/>

Figur 6-5 The simulation settings.

7 Examples

The following sections contains example of the methods implemented in Babar. The data sets used in the examples are available in the Help->Examples menu in Babar.

7.1 Example data sheet: Nine studies of different species of bats

This section describes how to create a data sheet of a data set that takes the form of statistics of studies (e.g. mean, standard deviation and sample size). This form of data is required in most computations in Babar, e.g. Bayesian updating and combining means and variances.

Each row in the data set describes observed statistics for a study (e.g. a study for a specific site, species, element or source). The statistics describing each study must be given at least the following values

Group classification IDs	A column of type Classification which identifies a group of studies (e.g. Element with values Cs, Ur,..)
Study classification IDs	A classification column identifying each study within the group (e.g. Site with values “Stockholm”, “Uppsala”...).
Measurement distribution type	Either Normal or Log Normal.
Mean and SD	Arithmetic mean and standard deviation (for a Normal measurement distribution type)
GM or GSD	Geometric mean and geometric standard deviation (for Log Normal measurement distribution type).
N	The number of samples for the study.

The following steps describe how to create a data sheet that takes the form of statistics.

- 1) Add a new data sheet to the project. Call the data sheet Nine Bats. Select Statistics as the column format template of the new sheet. Open the data sheet in the data editor view. The data sheet has the following columns: Element, Species, Classification, N,Mean,SD,GM,GSD, Measurement distribution. There are more columns than these, but these are the most important.
- 2) Open the column format editor (by right clicking the new data sheet and selecting Edit columns). Here, the columns and their corresponding types are listed. It can be seen that the first three columns are of type “Classification”. They can hold textual values which can be used to classify or group studies. N,Mean,SD,GM,GSD are types which uniquely identify the corresponding statistics of the study. The column Measurement distribution holds the distributions type if the measurement model. In this case, the columns Element and Species will be used. For this example, the default columns can be kept with no changes.
- 3) Go to the Data Editor view. Enter the following data for the columns Species,Element,N,GM,GSD and Measurement distribution.

Species	Element	N	GM	GSD	Measurement distribution
Brown long-eared bat	Cs-137	5	0.11	3.08	Log normal
Common Noctule Bat	Cs-137	20	0.09	2.68	Log normal

Daubenton's Bat	Cs-137	2	0.01	3.95	Log normal
Kuhl's Pipistrelle Bat	Cs-137	6	0.34	15.20	Log normal
Lesser Noctule Bat	Cs-137	5	0.06	1.62	Log normal
Nathusius' Pipistrelle	Cs-137	51	0.10	4.42	Log normal
Parti-coloured Bat	Cs-137	11	0.04	1.68	Log normal
Serotine bat	Cs-137	17	0.26	4.25	Log normal
Soprano Pipistrelle Bat	Cs-137	2	0.23	4.07	Log normal

7.2 Example data sheet: Random measurement values

This section describes how to create a data sheet of a data set that takes the form of raw data values. This form of data is required in Babar for fitting probability density functions to data. A column containing raw data values must be of type Value. For this kind of data, values representing different groups must be stored column wise.

The following steps describe how to create a data sheet to contain data values:

1. Create a new data sheet to the project. Select a name for the data sheet and "Raw data" as the column format.
2. Open the column format editor (by right clicking the data in the project view and select "Edit columns..."). There are 20 columns of type Value as default. Rename the first five as follows (by clicking the first column of the table): Normal, Log Normal Chi square, Weibull and Gamma.
3. Open the data sheet in the data editor view and enter the data from Tabell 7-1. The data are 15 samples drawn from the following distributions: Normal(mean=3,sd=1), Log normal(mu=3,sigma=4), Chi squared(3), Weibull(3,1) and Gamma(3,1).
4. Save the data sheet as FiveRandomVectors.

Tabell 7-1. Data generated from five different probability distributions.

Normal	Log normal	Log normal BDL	Chi squared	Weibull	Gamma
3.1832	46.5429	<2	3.1094	0.6660	6.9223
1.9702	8.2645	<2.7	0.9609	0.4067	4.3369
3.9492	22.2000	<3.5	5.7963	0.6886	1.0505
3.3071	11.6519	<4.2	1.5987	0.3182	3.2662
3.1352	27.2083	4.4542	8.5146	0.0120	11.3269
3.5152	11.0196	4.7157	4.4606	0.4949	2.5650
3.2614	32.7848	5.1464	1.9974	0.4214	2.3458
2.0585	42.0712	5.2927	2.1725	0.1069	5.9140
2.8377	111.2620	5.3015	0.7997	0.6432	5.7150
2.8539	16.5416	5.7138	1.9673	0.0560	1.1409
2.4680	2.3671	6.2077	4.4276	0.0892	6.3130
4.6821	8.6747	6.3192	3.4084	0.0700	4.7383
2.1243	77.8352	6.5527	2.9443	0.0353	2.2012
2.5162	6.8747	6.5690	0.8806	0.0403	1.5766

2.2880	52.5074	6.9295	2.8998	1.3865	1.1979
--------	---------	--------	--------	--------	--------

7.3 Example: Testing means and variances of species of bats

The following section uses the data set of nine species of bats defined above. The steps below describe how to perform tests on means and variances of the nine species of bats.

Because the studies of the bats have a Log normal measurement model and statistics given as geometric mean and geometric standard deviation, the means and variances are tested on logarithmic measurement scale. That is, tests are performed for $\ln(GM)$ and $\ln(GSD)$.

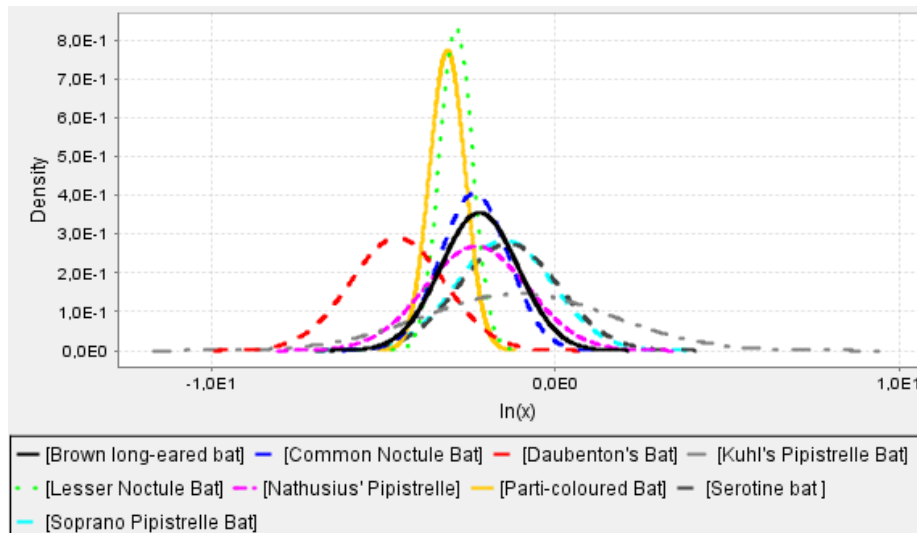
Plotting the observed distributions

- 1) In the project view, select the data set for the nine species
- 2) In the Analysis perspective, make sure all nine species are shown in the Analysis data table. If not, select all species and Element Cs-137 in the filter panel.
- 3) Go to the Analysis Chart view (section 5.3.2). Each species is represented by a normal probability density functions, derived from logarithmic mean $\mu = \ln(GM)$ and standard deviation $\sigma = \ln(GSD)$. The plot should look similar to Figur 7-1. Also select Species as the group by column in the filter. This will set the default label to the name of the species.

Testing means and variances

To test the mean and variances (μ and σ^2) perform the following steps:

1. In the Analysis view, select Element as the group by column in the filter. This will make sure the tests are performed between the studies with the same Element (in this example, all nine studies for Cs-137 are tested for equal variances). Select all nine species in the filter.
2. Go to the Analysis->Test mean and variances tab (5.2.1). Select Test of equal means and Test of equal variances and enter 0.05 as alpha. Press Test to start the tests.
3. The results of the tests are shown in the Simulation log window and should look similar to Figur 7-2. The p-value of the test of equal variances are $p < 0.01$ and indicates that the variances are indeed different (based on the chosen significance level 0.05). The test of means also report a low p-value ($p < 0.01$) but it should be kept in mind that the different variances violates the assumption of the ANOVA test.



Figur 7-1. Normal distributions showing the studies of nine species of bats.

16:17:29 CEST: Starting task Mean/Variance test...

Test group Cs-137

Equal means (ANOVA)

Source of Variation	SS	df	MS	F	Fcrit	P-value
Between studies	44,98	8	5,62	2,92	2,02	0,0054**
Within studies	212,00	110	1,93			
Total	256,97	118				

Equal variances (Bartlett's test)

Bartlett's test statistic	24,98
Chi2 critical value	15,51
Degrees of freedom	8
P-value	0,0016**

Means are different ($p < 0.05$) for batch Cs-137

Variances are different ($p < 0.05$) for batch Cs-137

16:17:29 CEST: Finished Mean/Variance test!

Figur 7-2. Output from the test of equal variances of nine species of bats.

7.4 Example: Combining means and variances of species of bats

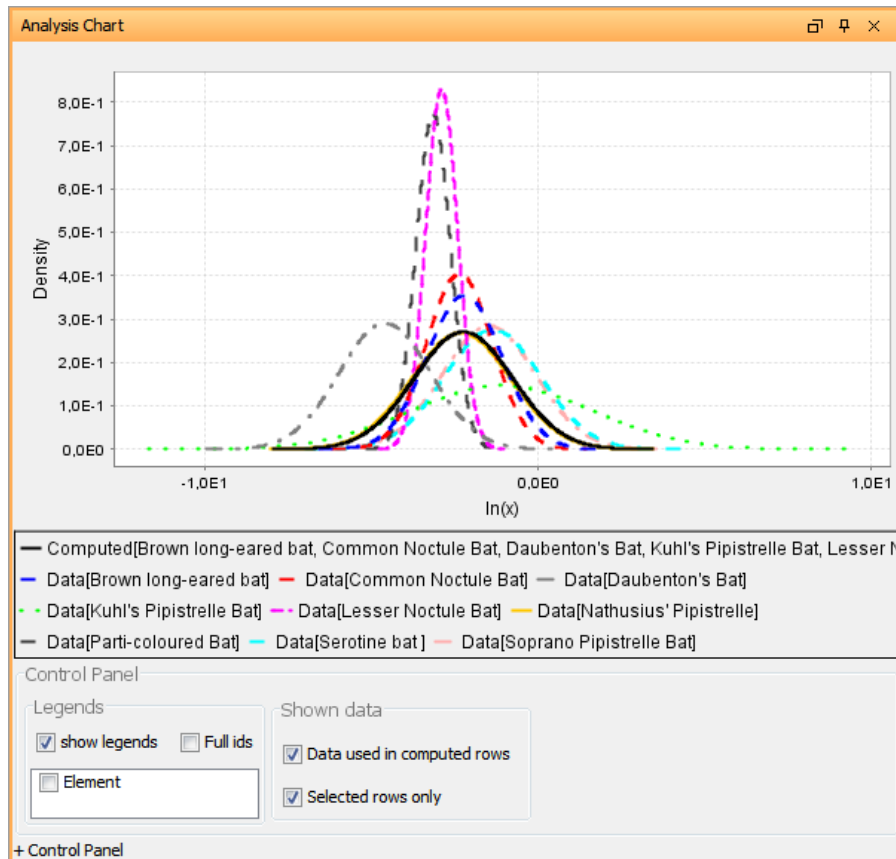
The steps below describe how to combine means and variances from eight of the species of bats from the data set defined above. The resulting statistics will then describe the collective data set based on the eight species. Because the studies of the bats have a Log normal measurement model and statistics given as geometric mean and geometric standard deviation, the computations are performed on logarithmic measurement scale. That is, the combination of means and variances are performed for the log transformed mean and variance, $\mu = \ln(GM)$ and $\sigma^2 = \ln(GSD)^2$.

1. In the Analysis view and the filter panel, select Element as the Group by column. This will instruct Babar to combine studies that have the same element. In the filter under column species, selecting all species *except Daubenton's Bat*.
2. In the Analysis->Pooling, select Pool means and variances and select include between study variance. This will instruct Babar to use the formulas for combined means and variances from section 2.2.1. Press Pool to compute the combined means and variances.

3. Save the resulting statistics to the Nine Bats data sheet: In the Result tab, Select “Nine Bats” data sheet. Then press Export to save the resulting row to the original data sheet. The value of the Species column of the exported row is stored as a comma separated list of all eight species combined. In the data editor view, change the Species value of the computed row to “CombinedExclDaubenton”. The data sheet should look similar to figure Figure 7-3. Note the value of in the column “Performed operations” that indicates that the last row is the result of a computation.
4. The resulting statistics can be plotted in the Analysis Chart (section 5.3.2) by selecting the result data sheet. As default, all rows are plotted (here there is only one row). Alternatively, select the row to be plotted and make sure “Shown data: is selected rows only” is selected in the Analysis chart control panel. The plot should look similar to Figur 7-4. As default, the PDFs corresponding to the data used for the computations is plotted together with the computed PDF(s). This can be turned on and off by selecting/deselecting “Shown data: Data used in computed rows” In the Analysis view control panel. The graph can be saved to an image file (.png) or copied to the clipboard from the context menu (opened by right clicking the chart).

	Species	Element	N	GM	GSD	Measurement d...	Performed operations (computed)
1	Brown long-eared bat	Cs-137	5	0.1099	3.078	Log normal	
2	Common Noctule Bat	Cs-137	20	0.0948	2.6838	Log normal	
3	Daubenton's Bat	Cs-137	2	0.0101	3.9506	Log normal	
4	Kuhl's Pipistrelle Bat	Cs-137	6	0.3395	15.1956	Log normal	
5	Lesser Noctule Bat	Cs-137	5	0.0557	1.6176	Log normal	
6	Nathusius' Pipistrelle	Cs-137	51	0.1003	4.4206	Log normal	
7	Parti-coloured Bat	Cs-137	11	0.0432	1.6756	Log normal	
8	Serotine bat	Cs-137	17	0.2566	4.2507	Log normal	
9	Soprano Pipistrelle Bat	Cs-137	2	0.2299	4.0705	Log normal	
10	CombinedExclDaubenton	Cs-137	117	0.1112	4.2621	Log normal	Pooled: Combined means and va...

Figur 7-3. The data for the nine species of bats and the computed statistics for the combined mean and variance of eight species (excluding Daubenton's Bat).



Figur 7-4. Plot of the combined means and variances in the Analysis chart view. The PDF corresponding to Computed[...], here solid black lines, corresponds to the combined mean and variance. The other PDFs show the data used for the computations.

7.5 Example: Bayesian updating of a population with Daubeton's bat

In this section, the combined mean and variance for the eight species excluding Daubenton's Bat (calculated in 7.4) will be seen as representing the population of bats and will be updated with data for Daubenton's bat. Two methods will be used: Conjugate updating and semi-conjugate updating.

Conjugate updating

In conjugate updating, the data sets will be combined treating them as exchangeable. The following steps describe how to perform the updating using the statistics for the population of bats as prior data and the statistics for Daubenton's bat as "observed data".

1. In the Analysis view and the filter panel, select the two species to combine under the species column. Also select Element as the group by column
2. In the Analysis->Direct updating select Species from the *Class column* list, select "CombinedExclDaubenton" as the prior and "Daubeton's Bat" as observed data. Select *Conjugate prior*.
3. Review the simulation settings. For Bayesian updating, all simulation settings are used but the most important is the number of iterations which should be at least 10 000. The Markov Chain settings can be left at their default values or set to *Burn-in: 500, Thinning factor: 1, Number of chains:3, Dispersion:0.2, Estimation method: Posterior median*. See 6.1.6 for details about the simulation settings.

4. Press Run to start the simulation from the posterior distribution. After successful simulation, result statistics (derived from the median of the marginal posterior distributions of the mean and variance) are shown in the Analysis view table (the result should look similar to the first row of Figur 7-5).
5. Save the computed row to the Nine Bats data sheet. In the Result tab, select the Nine Bats data sheet in the list and press Export.

Semi-Conjugate updating

In semi-conjugate updating, the population distribution (derived from eight of the bats) is interpreted as defining prior probabilities of the mean but is not used to provide any information of the variation within Daubenton’s Bat. The following steps describe how to perform the updating using the statistics for the population of bats as prior data and the statistics for Daubenton’s bat as “observed data”.

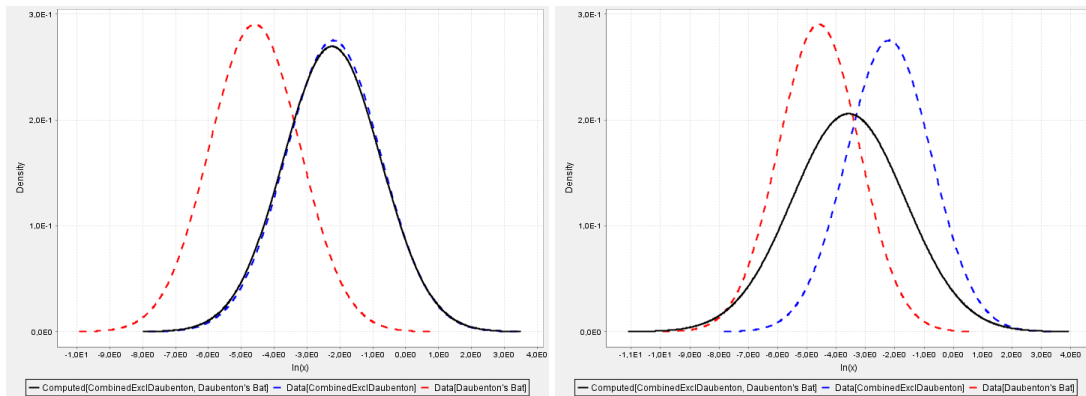
1. Follow the same steps as for Conjugate updating, but select *Semi-conjugate prior*. Select *Non-informative prior* as the value of the Variance estimation. This will consider the variance as uncertain and use a non-informative prior to estimate it its posterior distribution. Estimating it as a point estimate is highly unreliable since there is only two data points for Daubenton’s Bat.
2. Run the posterior simulation.
3. Save the computed row to the Nine Bats data sheet. In the Result tab, select the Nine Bats data sheet in the list and press Export. The resulting computed row should look similar to the second row in Figur 7-5).

The observed and posterior predicted statistics can be plotted in the Analysis Chart view (as normal distributions on log scale).

1. In the Analysis Chart view control panel, select “Data used in computed rows” and “Selected rows only”.
2. In the Analysis table showing the Nine Bat data sheet, select the row corresponding to the posterior from the updating of the conjugate prior. The result should look similar to the left chart of Figur 7-6. It is seen that the conjugate prior updating results in a posterior that is very close to the combined distribution of the eight bats. This is because only the number of data points is used to distinguish between the prior (the combined eight species based on 117 data points) and the data (Daubenton’s bat based on only two data points).
3. Select the row corresponding to the posterior from the updating of the semi-conjugate prior. The resulting chart should look similar to the right chart of Figur 7-6. The posterior predicted distribution of Daubeton’s bat is closer to the data for the Daubenton’s bat although updated to adapt to the prior information of the mean. The posterior variance is increased to account for updated mean.

Species	Element	N	GM	GSD	Measurement ...	Performed operations (computed)
CombinedExcdDaubenton, Daubenton's Bat	Cs-137		0.1068	4.3919	Log normal	Conjugate prior
CombinedExcdDaubenton, Daubenton's Bat	Cs-137		0.028	6.9368	Log normal	Semi-conjugate prior

Figur 7-5. Posterior predicted GM and GSD of population distribution combined with Daubenton’s bat using Bayesian updating using a Conjugate prior (first row) and Semi-conjugate prior (second row).



Figur 7-6. Normal distributions with parameters $\ln(GM)$ and $\ln(GSD)$ derived from statistics of the Daubenton's bat, the combined eight bats and the posterior distribution for Conjugate prior (left) and semi-conjugate prior (right).

Inspecting posterior distributions and checking convergence

Statistics and convergence measures of the posterior distributions can be inspected in the Simulation output statistics view. The view shows three simulation outputs, $\mu[Cs-137]$ and $\sigma^2[Cs-137]$ which represents the mean and variance of the logarithmic measurement model. The third output $Pred[Cs-137]$ are samples from the predicted distribution. The convergence statistic R (the Gelman Rubin estimate of potential scale reduction) should at least below 1.001 in this example, indicating that the “width” of the posterior distributions can be reduced by at most 0.1% if the simulation were to continue. For the posterior distributions from the semi-conjugate method, the samples for $Pred[Cs-137]$ can yield extremely high values and the statistics might even be infinite. The output is however not used for inferences if the Estimation method (in the simulation settings) is set to Posterior medians. For this choice of estimation method it is only the posterior outputs μ and σ^2 that are of interest.

7.6 Example: Hierarchical updating of eight species of bats

The following section describes how to perform hierarchical updating of eight species bats using the data set defined in section 7.1. One of the nine species of the data set, Kuhl's Pipistrelle has an extreme variance ($GSD=15$) and is excluded from the estimation. It will be assumed that the eight bats are exchangeable, in that there is no information available that distinguish the species. Furthermore, it is assumed that within species variance can be estimated as one common variance (using the assumption of homogeneous variances).

1. Select the data sheet Nine Bats in the projects view.
2. In the Analysis view and the Analysis->Filter tab, select the all species of bats except Kuhl's Pipistrelle. Select Species as the group by column (The group by column for hierarchical updating is used to instruct Babar how to distinguish between hierarchical cases or groups of measurements). In this case, we have eight cases (species) that should be simultaneously estimated.
3. In the Analysis->Hierarchical updating tab, select *Variance Estimation: Homogeneous with non-informative prior*.
4. Review the Simulation settings. Make sure the number of iterations is set to at least 100 000 and Estimation method is *Posterior median*. Start the posterior simulation by pressing *Run*.
5. After successful simulation, the simulation outputs are shown in the Simulation output statistics view. Make sure all values of R (the Gelman Rubin convergence statistic) is below 1.001. This indicates that the width of the posterior distributions can be decreased by an approximate maximum of 0.1% if the simulation where to continue.

6. Statistics derived from the medians of the posterior distributions for the eight species are shown in the Analysis view table.

Species	Element	GM	GSD	Measurement distribution	Performed operations (computed)
Brown long-eared bat	Cs-137	0.1065	4.0715	Log normal	Hierarchical updating; Variance estimation: Homogeneous
Common Noctule Bat	Cs-137	0.0968	4.0715	Log normal	Hierarchical updating; Variance estimation: Homogeneous
Daubenton's Bat	Cs-137	0.047	4.0715	Log normal	Hierarchical updating; Variance estimation: Homogeneous
Kuhl's Pipistrelle Bat	Cs-137	0.204	4.0715	Log normal	Hierarchical updating; Variance estimation: Homogeneous
Lesser Noctule Bat	Cs-137	0.075	4.0715	Log normal	Hierarchical updating; Variance estimation: Homogeneous
Nathusius' Pipistrelle	Cs-137	0.1008	4.0715	Log normal	Hierarchical updating; Variance estimation: Homogeneous
Parti-coloured Bat	Cs-137	0.0563	4.0715	Log normal	Hierarchical updating; Variance estimation: Homogeneous
Serotine bat	Cs-137	0.2088	4.0715	Log normal	Hierarchical updating; Variance estimation: Homogeneous
Soprano Pipistrelle Bat	Cs-137	0.1323	4.0715	Log normal	Hierarchical updating; Variance estimation: Homogeneous

Inspecting posterior distributions and checking convergence

Statistics and convergence measures of the posterior distributions can be inspected in the Simulation output statistics view (Figur 7-7). The posterior outputs corresponding to mean parameters (labeled mu) is all on log scale, meaning that they have to be transformed if to be represented on the same scale as the measurements. The parameter mu.pred is predicted (log) means from the population distribution and can be used to represent possible values of a new species of bat considered to belong to the same family. The convergence statistic R (the Gelman Rubin estimate of potential scale reduction) of all posterior outputs are below 1.001 indicating good sufficient convergence.

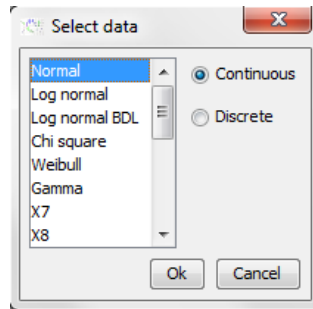
Output	N	R	MCSE	Mean	SD	GM	GSD	2.5% percentile	5% percentile
mu[Brown long-eared bat]	28 500	1.0000	0.0027	-2.2886	0.4594			-3.1956	-3.0374
mu[Common Noctule Bat]	28 500	1.0001	0.0016	-2.3622	0.2674			-2.8939	-2.8033
mu[Daubenton's Bat]	28 500	1.0005	0.0046	-3.2876	0.7778			-4.9859	-4.6942
mu[Lesser Noctule Bat]	28 500	1.0003	0.0028	-2.6849	0.4790			-3.6820	-3.5109
mu[Nathusius' Pipistrelle]	28 500	1.0000	0.0010	-2.3052	0.1772			-2.6528	-2.5972
mu[Parti-coloured Bat]	28 500	1.0006	0.0023	-2.9374	0.3835			-3.6980	-3.5772
mu[Serotine bat]	28 500	1.0002	0.0020	-1.5773	0.3372			-2.2582	-2.1471
mu[Soprano Pipistrelle Bat]	28 500	1.0000	0.0038	-2.0308	0.6335			-3.1973	-3.0007
mu.pop	28 500	1.0001	0.0023	-2.4317	0.3824			-3.2429	-3.0619
mu.pred	28 500	1.0000	0.0060	-2.4347	1.0109			-4.5666	-4.0499
sigma.pop^2	28 500	1.0001	0.0069	0.8719	1.1700	0.4913	3.3626	0.0338	0.0689
sigma^2	28 500	1.0000	0.0014	1.7216	0.2448	1.7047	1.1503	1.3079	1.3614

Figur 7-7. Summary and convergence statistics for the hierarchical simulation outputs. The top eight outputs are the posterior means of each species (on log scale). The last output is the posterior of the common measurement variance.

7.7 Example: Distribution fitting of observed measurements

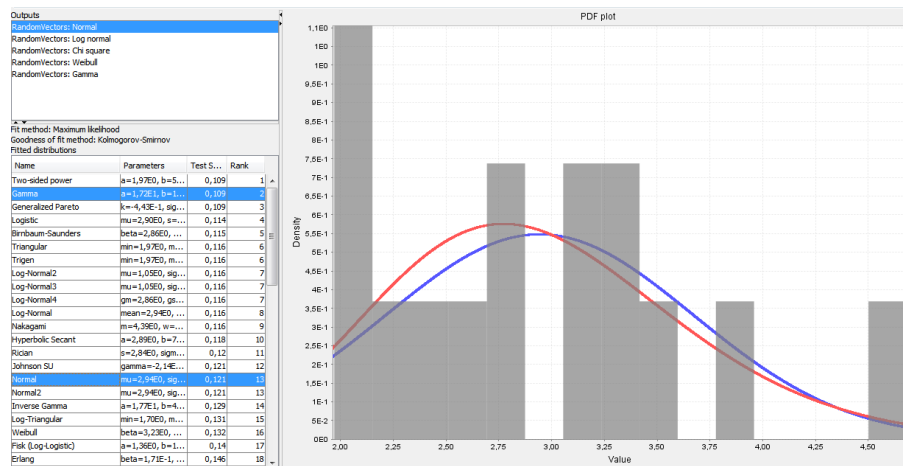
The following steps describe how to perform fit probability distributions to data values. The example uses the data for the six random vectors defined in section 7.2.

1. Open and select the data sheet SixRandomVectors created in section 7.2.
2. Go to the Distribution Fitting perspective. Press "Load Samples..." and select the column in the data sheet containing the data to fit (Figur 7-8). Select all columns one by one. Keep "Continuous" selected as the samples are continuous measurement (i.e. not discrete).



Figur 7-8. Dialog to select samples to load into the distribution fitting perspective.

3. In the list Outputs. Select Normal and press Fit. This will start the fitting of available continuous distributions. When finished. The result of each fit is shown in a table in order of best fit. For the 15 sampled values, a normal distribution is fitted with parameters $\mu = 2.94, \sigma = 0.73$. The ranking of the normal distribution is low (rank=13 based on the Kolmogorov Smirnov statistic).
4. The summary statistics of the sample and one selected fitted distribution can be compared in the Fit Statistics view. In this view, the Number of samples, mean,SD,GM,GSD,



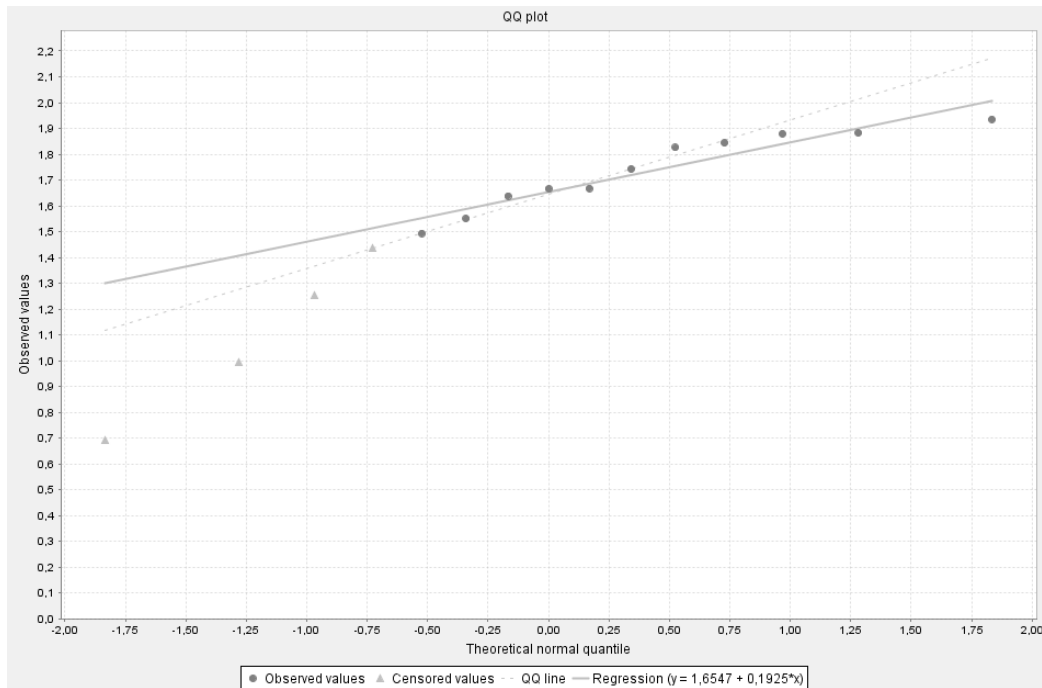
Figur 7-9. The 15 samples from a normal(3,1) distribution shown as a histogram and the result of fitted distributions. In the figure, the Gamma(a=1.7,b=0.17), with rank 2 and Normal(2.94,0.73), with rank=13 is plotted over the histogram.

Fitting distributions to values below detection limit (BDL)

The column Log normal BDL contains values below detection limit. These are specified in the data sheet as by prepending the value with a “less than sign, <”. A value such as “<0.1” will thus be interpreted by Babar as being a value less than 0.1. Vectors containing at least one value below detection limit cannot be fitted with the usual methods. The method used by Babar is instead based on a technique based on the empirical cumulative probabilities explained in (section 2.5.2). This method can fit only normal or log normal distributions.

1. Load the samples in the column Log normal BDF into the distribution fitting perspective. Select the values in the Output list.
2. Click Fit to fit normal and/or log normal distributions to the values. The resulting distributions will be listed in the Fit result list. Ranking is not possible for this method. Choice of distribution (normal or log normal) should be based on theoretical knowledge of the observed variable, concentrations for example are often known to be log normally distributed rather than normal.

- Choose QQ-plot as the plot type, select $\ln(x)$ to plot logarithmic values and select “Show regression line”. The figure should look similar to Figur 7-10.



Figur 7-10. A quantile plot of a vector of log normal samples (here log-transformed), where four values are below detection limit (BDL), and the corresponding theoretical quantiles (from a $N(0,1)$ distribution). A regression line is fitted to the quantiles and observed values. The intercept and slope is used as estimates to the mean and standard deviation of all (log transformed) values.

7.8 Examples: Weighted resampling

Weighted resampling from are performed by following the following steps.

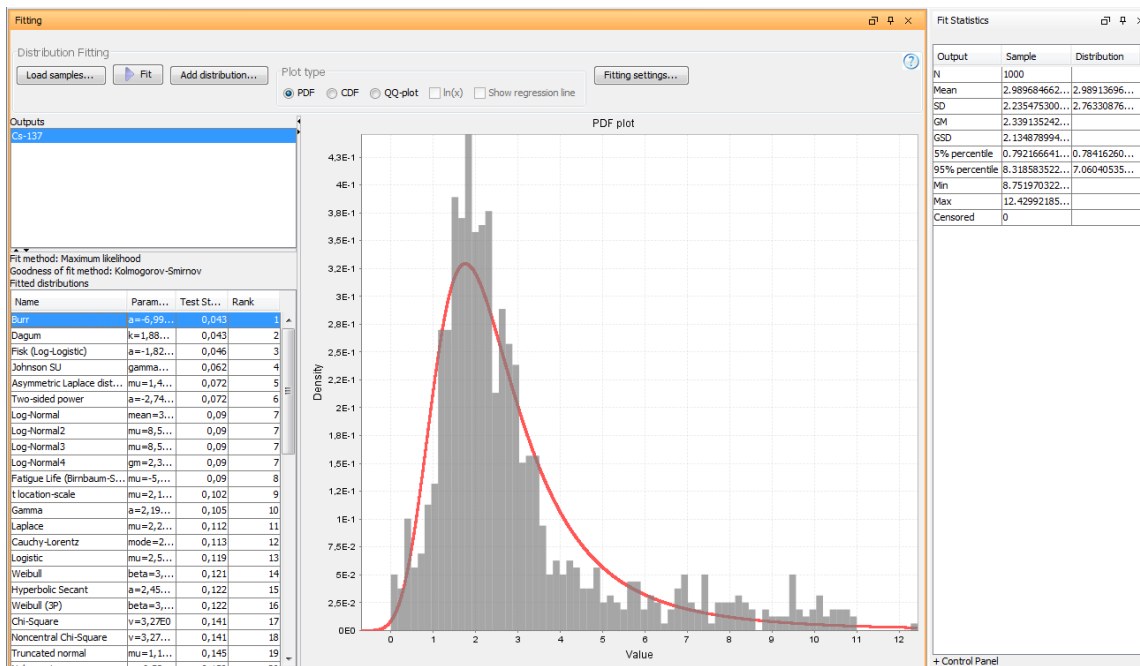
- Four different probability distributions are available for a certain variable (Figur 7-11). Each distribution has been given a number N to quantify the weight of certainty to the particular PDF relative the other PDFs. In this example, Expert 2 is given most certainty (double that of Export 1), literature A and B are given equal credibility but only half that of Export 1.

	Variable	Source	N	Distribution
1	Cs-137	Expert 1	10	$\text{logn}(2.0, 0.5)$
2	Cs-137	Expert 2	20	$\text{logn}(2.6, 1.3)$
3	Cs-137	Literature A	5	$\text{triang}(0.0, 6.0, 2.5)$
4	Cs-137	Literature B	5	$\text{exp}(2.0)$

Figur 7-11. Data used for sampling/weighted resampling of probability distributions

- In the Analysis->filter tab, select –none- or “Variable” as group by column. This will tell Babar to sample from all four distributions given their relative weights given by N .
- In the Analysis->Resampling tab, go to simulation settings and enter 1000 as the number of samples (a moderate number is recommended since too many samples will take a long time to fit numerically). Press Generate Samples to start to sample from the PDFs. After successful simulation, the samples are sent to the Distribution Fitting perspective.

4. In the Distribution Fitting Perspective, select Cs-137 in the list of Outputs. The samples are plotted as a histogram. Click "Fit" to fit all distributions to the samples. When finished, the list of ranked distributions are shown in the result list. The view should look similar to Figur 7-12. The resulting histogram is formed by the four PDFs but is dominated by the PDFs with largest weights. Summary statistics (Mean,SD,GM,GSD,min,max and lower and upper percentiles) are shown for the resulting samples and for one selected fitted distribution. The figure below shows the result for the highest ranked distribution (a four parameter Burr distribution).
- 5.



Figur 7-12. Distributions fitted to 1000 samples from the weighted resampling.

8 References

Burmaster D E, Hull D A, 1997. Using Log Normal Distributions and Log Normal Probability Plots in Probabilistic Risk Assessments, Human and Ecological Risk Assessment, Volume 3, Number 2, pp 235–255.

Casella, G., George E., (1992). Explaining the Gibbs Sampler. The American Statistician 46, (3), 167-174.

Gamerman D., Lopes F. H., 2006. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, second edition (Chapman & Hall/CRC).

Gelman, A. 2006. *Prior distributions for variance parameters in hierarchical models, Bayesian Analysis. 2006/.*

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2004. Bayesian Data Analysis. Chapman & Hall, 2nd edition, 2004. ISBN 1-58488-388-X.

Gelman, A., Hill, J., 2007. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press. ISBN 0-521-86706-1.

Morris, C. N., 1983. Parametric empirical Bayes inference: theory and applications., Journal of the American Statistical Association, 78, 47-65