DECISION: Getting started.

Preface to DECISION 3.6

The purpose of DECISION is to help making decisions based on data. Using DECISION may reduce the number of computer runs, estimate the effects of the different input variables, and finally determine the uncertainty in the response variables given the distribution of the input variables. DECISION 3.6 is documented in

- **DECISION: Getting started**,
- **DECISION: Theory, Methods and Examples**,
- **DECISION: Users reference manual**.

DECISION 3.6 is coded using C++ and is object oriented. Version 3.6 is the last regular version of the 'Kapof' part of the project. There will be new releases resulting from extensions and maintenance. Moreover, history matching relates closely to the DECISION project and may be viewed as a separate module financed by Norsk Hydro and Statoil, documented in the NR/SAND note 'Automatic History Matching, Part 2'.

The developers of DECISION gratefully acknowledge help from Eivind Damsleth and Knut Hollund of Norsk Hydro, Adolfo Henriquez, Håkon B. Eilertsen, Svein Tore Opdal and Kjetil Skrettingland of Statoil and Geir Tyberø of Conoco. Finally, we thank Alan Miller and Nam-Ku Nguyen who provided the code to handle D–optimality — we have modified their code and are responsible for errors.

Preface to DECISION/History Matching v 3.7.1.

This version of the manual covers new probability distributions that have been added to DECISION lately. Moreover, flexible means of describing dependence between variables have been introduced. A new chapter that aims at assisting the user when selecting distribution is included. See Chapter 4 [Distributions], page 27.

Finally, documentation on the History Matching module is included. Since not all companies with DECISION installations have chosen to buy the history matching module, that section may be irrelevant to some readers.

1 Introduction

The program is named DECISION to indicate that its ultimate aim is to aid in decision making. DECISION is designated to perform sensitivity studies of computer intensive computations. The effect of using DECISION is to reduce the number of runs and increase the precision of estimates. A typical application will be based on output from a large reservoir simulator like ECLIPSE. A relation between the response variable (say 'Production after 4 years') and the input variables (Lobes, STOOIP,...) is estimated (generally there may be several response variables). This relation is used as a *surrogate simulator*: The values of the input variables are sampled from assigned probability distributions to produce a large number of Monte Carlo simulated values for the response variable. Based on these values the distribution of the response values may be plotted and characteristic values like mean and standard deviation calculated. In the case of several response variables a production profile with uncertainty may be produced.

Another application of DECISION is in history matching and other optimization problems. Optimization is done using the estimated relation between input and response variables.

The items below are the main modules of DECISION. The numbering indicates the standard order used.

1. Variables
2. Design
3. Analysis
4. (History Matching / Optimization)
5. Simulation

Calculations are performed in modules 2-5 while the first module is used to define and code variables. Plotting is performed using the '`Graph`' options of the various windows. The purpose of the '`Design`' module is twofold: first the candidate set is generated, then the best runs are selected from the candidate set. The '`Analysis`' module analyzes the data deriving a predictor, i.e., the optimal relation between the input variables and the response. Currently there are three more or less related methods that may be used to analyze the data: regression, automatic regression and kriging. The '`History Matching / Optimization`' module is can be used to search for input values that will make the response match observed historical values, or it can be used to search for input values that will minimize/maximize a response. The purpose of the '`Simulation`' module is mainly to provide a number of values for the response variable: distributions are assigned to the input variables and Monte Carlo simulation is performed. In addition it is possible to perform Monte Carlo simulation based on a user specified equation.

In history matching and optimization, the user will go from the analysis module to the History Matching module, to perform optimization using the response surface.

The purpose of the present manual is to explain the main features of DECISION. It is advantageous to have access to DECISION while reading the present manual. To get a more complete understanding of DECISION including possibilities not covered in this introduction manual, the reader will have to consult '`DECISION: Theory, Methods and Examples`' or the '`DECISION: Reference manual`'. Newcomers to X/Motif user interface are recommended to read Appendix E of the latter manual to learn how to handle menus and textfields and so on.

Also note that this manual is available both as a printed document and through DECISION's on-line help system.

Some very rough plots are included in some sections. They look so rough because they are typeset in a way that can be used both in the printed and the on-line version of the manual.

2 Getting started

This chapter provides a detailed description of how to use the main features of DECISION intended for the beginner or the inexperienced user. We will explain the features of the Control Panel (or Project window), Variables, Design, Analysis, Simulation and History Matching/Optimization windows as well as the Graph options.

2.1 The Control Panel window

When you start DECISION a window we call the Control Panel or Project window appears on the screen. The contents of this window is from top to bottom:

- A menu bar with buttons labeled '`Project`' and '`Help`'.
- A set of buttons labeled '`Variables`','`Design`','`Analysis`', '`Simulation`' and '`History`'.
- The text '`New.prj`'.
- A text field reading '`Enter project description here!`'. This text is not used by DECISION.

A new project is started selecting '`New`' in the '`Project`' menu. Note however that by default a project is started with some default values DECISION.

We will start by defining some variables. This is done in the Variables window: Point the mouse cursor at the '`Variables`' button and click the left mouse button. A new window named '`Variables`' will appear.

2.2 The Variables window

Before we can do any work on the new project we must define the input and response variables we will use.

The window has a menu bar with menus labeled '`Response`', '`Input`', '`Distribution`', and '`Help`'. The window itself consists of two parts called '`Response Variables`' and '`Input Variables`'. Both the '`Response Variables`' list and the '`Input Variables`' list contain one variable initially.

The pull down menus '`Response`' and '`Input`' are used in order to enter, delete or copy a variable in the corresponding list. In addition there is a report option in '`Response`'. If this is chosen, a file selection dialogue box pops up prompting for a filename. A description of the variables is written to the indicated file. The 'New' option creates of the pull down menus 'Response' and 'Input' create a new variable identically to the initial line. The 'Copy' option makes a copy of and the 'Delete' option deletes a variable in the table. The last two options are only active when a variable in the table has been selected by a mouse click on the right hand side button.

The pull down menu 'Distribution' is used in order to change the distribution type of a selected variable in the 'Input Variables' list. It also inserts default parameter values.

The default values of the variables in the two variables lists often need to be edited. To do this, click on the item with the right mouse button, make the changes, and hit RET on the keyboard ("return") when finished.

Each variable in the 'Response Variables' list consists of a short name (default names for new variables are 'Y0', 'Y1' etc.) and description (default is 'New Variable'). The next field 'Truncation' is only relevant for simulation. Leaving the field unchanged (i.e. the default specification (NaN,NaN) (Not a Number)) is the normal usage. The user has however the option to truncate the simulated values (described later on) to an interval. To be specific, suppose (0,10) is entered. Then all simulated values of the corresponding response variable will be constrained to this interval. If, say a value of 10.9 is drawn, the value is moved to 10. If there are to many moves (more than 5% of the number of simulations), the user is warned. The user may similarly perform one sided truncation, by entering say (,10) or (0,). This option should be used with care; one should not uncritically discard outlying simulations. The next column 'Series' is only relevant as far as plotting goes. The user has the opportunity to order the variable on some time scale. To be specific, create 4 response variables 'Y0,Y1,Y2,Y3'. If the last column is made to read 't(0),t(1),s(1),s(0)', it is signalled that 'Y0' and 'Y1' belongs to the same time series and that 'Y0' precedes 'Y1'. The last 2 variables belong to a different time series and 'Y2' succeeds 'Y3'. This allows the user to (later on) plot the variables on the proper time scale.

Each variable in the 'Input Variables' list also contains a short name (default names for new variables starts with 'X') and a description. In addition the variable is defined on three levels (default values:0,-1,1). Note that the base case (here 0) is given first. The last item is the distribution type. The default distribution is a house distribution. Distribution options are:

- Normal(mean, std. deviation)
- Truncated Normal(mean, std. deviation, lower limit, upper limit)
- Lognormal(mean log, std. deviation log)
- Truncated Lognormal(mean log., std. deviation log., lower limit, upper limit)
- Beta(shape parameter a, shape parameter b)

- Gamma(shape parameter t, shape parameter p, left end-point x0)
- House((x1, y1), (x2, y2), (x3, y3))
- Discrete((value, chance), . . . , (value, chance))
- Deterministic(value)
- Histogram((value, Percentile), . . . , (value, Percentile))

Consult the manual '`DECISION: Theory, Methods and Examples`' for further information about these distributions.

Now you are able to edit the variable lists. Try this: Click on the description field of the response variable, change its description to '`Response 0`', and change the input variable so that it has description '`Input 0`'. Changes in the distribution field are done in two steps. If you want to change the distribution type, this is done by selecting from the pull down menu. Parameter changes are inserted by editing directly in the distribution field.To add another input variable with the same name, description, levels and distributions, click on X0 (note that the X0 line is highlighted) and select '`Copy`' in the '`Input`' menu. A new line which is a copy of the first appears in the bottom of the list. Now edit its name and description, changing them to '`X1`', '`Input 1`'. (An alternative to copy is '`Add Input`' in the '`Variables`' menu. A new line with default values in all fields will appear in the '`Input Variable`' list.)

If you want to follow the rest of this introductory example, make sure that you have defined two input variables ('`X0`' and '`X1`') and one response variable ('`Y0`'). The next step will be to set up a list of experiments that should be run to obtain maximum information on how the response vary as functions of the input variables. Select the Design window by clicking the '`Design`' button in the Control Panel.

2.3 Introduction to plotting

Plotting is activated from the different windows and is currently possible for the Design, Analysis, Simulation and History Matching windows. The main features of the various windows coincide. Note that virtually no plotting in the analysis and simulation windows are possible prior to actually running the analysis or simulation. Full details on plotting are provided in Chapter 8 of DECISION: Reference manual, where documentation written December 1993 by Christian Buchholz, TSC a/s, is included.

2.4 The Design window

This window has a menu bar with buttons labeled '`Design`', '`Candidates`', '`Experiments`', '`Graphs`' and '`Help`'. The window itself is divided into 3 parts. The first part contains two text fields, called '`Selection`' and '`Candidates`'. The middle part of the window is named '`Selected Design`' and the lower part of the window is named '`Candidate Set`'. The use and meaning of these fields will be explained below.

The purpose of the design window is to set up a list of experiments that are the optimal experiments to run for exploring the relationship between the input and response variables. By an experiment we mean a set of input variables and the corresponding response values. The response value will be '`NaN`' if the experiment has not yet been run.

Simplified the setting up of a design is done in two steps: First a list containing all possible experiments is generated. This list is called the '`Candidate Set`'.

The next step is to select from the candidate set the optimal experiments. This set of experiments will be called the '`Selected Design`'.

Finally the selected experiments are run by some external program, and the calculated values for response variables are edited or read into into the '`Selected Design`' window.

2.5 Candidate set

The '`Candidates`' pull down menu contains several subfields. Before selecting the '`Generate`' button which actually starts the generation of the candidate set, you may choose between two different methods to be used. The '`Type`' field contains the selection buttons '`Rejection`' and '`MonteCarlo`' explained below. The selected method is shown in the '`Candidates`' field together with the default parameter(s). The parameter value(s) may be edited by user.

The rejection method is chosen by default, and in our case with two input variables the parameter is 2. The parameter value sets the maximum number of input variables that may differ from base value. In general, the user is well advised to select '`Rejection`' with the mouse button (if this option is desired) since this will give a sensible default value for the parameter. The default value ensures that no more than 1000 candidates will be generated

The MonteCarlo method requires two parameters: the number of experiments to be generated and a seed number. The values for an input variable are chosen randomly; each level of the variable is assigned the same probability. This may lead to duplicate experiments; duplicates are, however, removed in the selection procedure described below. The seed number is set by default or by the user in order to be able to regenerate the same sequence of experiments.

The '`Edit`' field of the '`Candidates`' pull down menu lets the user add or delete candidates. New candidates are created by making a copy of an existing one and then edit on it directly in the '`Candidate set`' table.

In order to force a candidate to be selected, enter a value in the response variable field. This may be a dummy value or the real value obtained by running the actual experiment. The candidates with a response value differing from '`NaN`', will also remain in the Candidate set table when a new set is generated. In many cases it may be desired to select experiments in several steps. Sometimes, only a few experiments are performed and analyzed. Based on the analysis, new experiments are selected. Further examples are provided later on.

Using the default setting in the '`Candidates`' field, select the '`Generate`' option of '`Candidates`' pull down menu in order to generate a candidate set. The candidates are shown in the '`Candidate Set`' field in the bottom of the Design window. In this case the candidate set consists of 9 possible experiments. Each experiment is given an identification number in the first column. You may need to use the scrollbar to the right of the list to see all entries.

The '`Experiments`' is used to load, save and run experiments; the last option is not in use as of now. It is only relevant if DECISION is combined with an external program. To save an experiment to file, the user selects the appropriate experiment and merely enters '`save`'. The file name provided has the suffix '`exp`' by default. Typically, the user will save an experiment from selected design. Once the external program has been run, the values of the response variables are entered in the file. The experiment is then loaded into DECISION using the load option. If the input variables are changed, the experiment id should be removed from the file.

The '`Report`' option of the '`Design`' pull down menu is used to get output of the design window to a file.

2.6 Selected design

The '`Design`' pull down menu contains several subfields. The selection function in the '`Selection`' field must be defined to begin with. This function defines the number of candidates to be selected and the anticipated relation between the input variables and the response variable(s). The default selection rule is obtained by pushing the '`Dopt`' button in the '`Design`' pull down menu.

The selected function is shown in the '`Selection`' field: '`Dopt(7,(1,X0,X1,X0*X0,X1*X1,X0*X1))`'. The default number of experiments is one more than the number of terms in the equation. Suppose we want to run all 9 experiments in the candidate set, and that we want to use the equation

```
Y0 = b0 + b1 * X0 + b2 * X1 + b3 * X0*X1 + b4 * X0*X0 + b5 * X1*X1
```

Then the selection rule should be edited to

```
Selection: Dopt(9,(1,X0,X1,X0*X1,X0*X0,X1*X1))
```

Observe that the final equation based on the analysis described below may differ from the one above: The above equation is typically assigned before all experiments are run. The final, estimated equation will typically differ.

'`Dopt`' means that D-optimality will be used to select experiments (see the '`DECISION: Theory, Methods and Examples`' manual for an explanation of D-optimality), '`9`' is the number of experiments to be included, and '`(1,X0,X1,X0*X1,X0*X0,X1*X1)`' are the terms in the initial equation. Note that since the selection box has enough space for several lines of text, `RET` is used as '`Carriage Return`'. To finish editing this box, use `ctrl-RET` (that is, hold down the control key on the keyboard while pressing "Return". Generally, this way of entering input will always work.).

To perform selection, click on '`Select(Candidates)`' in the '`Design`' menu. This will generate a list with 9 experiments in the Selected Design field. (Note that since all 9 experiments are requested the equation is irrelevant for this example. The reader may want to experiment with differing input).

After one selection has been made, the user may create a new '`Candidate set`' and apply the '`Select(All)`' button in the '`Design`' menu. In this case the new selection is made from both the existing selection and the new '`Candidate set`'.

When the experiments have been run, response values can be entered. In this example the results from the experiments are:

```
Y0 X0 X1
50  0  0
40 -1  0
58  1  0
32  0 -1
21 -1 -1
40  1 -1
70  0  1
58 -1  1
86  1  1
```

The feature of marking experiments in the '`Candidate Set`' (by entering values for the response variables) may be utilized when there are many input variables. In this case a combination of the '`Rejection`' method and the '`MonteCarlo`' method may be a good choice. A one-at-a-time plan may be generated using '`Rejection(1)`'. The experiments may be marked. Then, additional values may be generated by the '`MonteCarlo`' to secure a spread of the experiments.

There are some limited plotting possibilities (needs) in this window. The user may produce crossplots of requested input variables and some projections.

Before you can proceed with analysis and simulation you always have to enter at least as many response values as there are terms in the equation. When you have done this, open the Analysis window by clicking on the '`Analysis`' button.

2.7 The Analysis window

The '`Analysis`' window has menu bars with buttons labeled '`Analysis`', '`Calculator`', '`Model`', '`Graphs`' and '`Help`'. The objective to analyze the data. In the upper part of the window you find a response variables list, where name, model type and the quality of the model is given for each response variable reading

```
Y0  RegressionModel(3)    NaN
```

which means that Y0 will be analyzed with a regression model with 3 terms. The terms to be included in the regression model are shown in the Model Template field. The quality of the model is '`NaN`' since no analysis has been performed yet.

The only input is the equation which is shown in the '`ModelTemplate`' field. To start with the default model, select '`Default`' in the '`Model`' menu. The Model Template is

```
RegressionModel = (1, X0, X1, X0*X1, X0*X0, X1*X1)
```

Select '`Selected`' in the '`Analysis`' menu. This will cause the regression analysis to be performed for the highlighted response variable. If several response variables are defined, the '`All`' option of '`Analysis`' will perform analysis for all response variables. The most important results are:

- The predictor.

  The relation between the response variable Y0 and the input variables X0 and X1 is estimated and the resulting predictor is displayed (you may have to scroll to see the entire predictor):

  ```
  49.4444*1 + 10.8333*X0 + 20.1667*X1 + 2.25*X0*X1
               -0.166667*X0*X0 + 1.83333*X1*X1
  ```

  This equation is called the predictor because it can be used to predict Y0 when X0 and X1 are known.

- Standard Deviation

  ```
  Residual Standard Deviation = 2.22569
  Cross Validated Standard Deviation = 4.44568
  ```

  The residual standard deviation and the cross validated standard deviation listed above are defined precisely in the '`DECISION: Theory, Methods and Examples`' manual. We will be content to be present an informal description here. The residual of observation no. i is defined as the difference between observation i and the corresponding predicted value. To compute the residual standard deviation, sum the squared residuals and standardize and complete by taking the square root.

  Cross validation consists of setting aside parts of the data, in this case only one observation, and predicting that data from what remains. The cross validated standard deviation is computed as described above except that the predicted value of observation no. i is based on a model where observation i was not used to derive the model. The advantage of the cross validated estimate is that it punishes over-fitting. The cross validated estimate is therefore well suited for comparing different regression models. A different point of view is that by using cross validation some of the model uncertainty is accounted for. The cross validated estimate is usually larger than the traditional estimate.

- Warnings

  In our case only one type of warning pertains and the user is given the following (not unusual message):

  '`*Warning* 2 experiments are outliers. (number 4,10)`'

- Also note that the Quality field has now been filled. For regression models, multiple R–square is used as measure of quality. Multiple R–square is 0.995336 indicating that approximately 99.5 % of the variation in the data is accounted for by the model; this indicates a good fit. A word of caution is due: there are many terms (6) compared to the number of observations (9) and so one should not be impressed by the the high value of the multiple R–square. A more reasonable model will be derived shortly.

- A closer look at the terms follows; the standard errors of the parameter estimates are given as well as further details.

Term	coef	std.err	p.value
1	49.44	1.66	0.00
X0	10.83	0.91	0.00
X1	20.17	0.91	0.00
X0*X1	2.25	1.11	0.14
X0*X0	-0.17	1.57	0.92
X1*X1	1.83	1.57	0.33

The p–value of the X0 term is 0.00 implying that this term is important; the higher the p–value, the less important the corresponding term.

- The observations.

Id	Dep Var	Predict Value	Residual	Student Residual	Hat	Dffits
2	50.00	49.44	0.56	0.31	0.56	0.35
3	40.00	38.44	1.56	1.08	0.56	1.20
4	58.00	60.11	-2.11	-2.04	0.56	-2.28
5	32.00	31.11	0.89	0.52	0.56	0.58
6	21.00	22.36	-1.36	-1.89	0.81	-3.85
7	40.00	39.53	0.47	0.41	0.81	0.83
8	70.00	71.44	-1.44	-0.96	0.56	-1.07
9	58.00	58.19	-0.19	-0.16	0.81	-0.33
10	86.00	84.36	1.64	5.14	0.81	10.45

In the lower part of the window details pertaining to each observation are provided. The meaning of the first few columns are immediate whereas the last three are more or less expert output. These are explained in the manual '`DECISION: Theory, Methods and Examples`'.

- Correlation matrix

The correlations between the coefficient estimates are given. (You have to scroll to see the correlation matrix.) It is nice if the off-diagonal elements are close to zero. In this case the different terms of the predictor may be interpreted separately. In technical terms the design is orthogonal provided the off-diagonal elements vanish.

p-values are numbers between 0 and 1. What p–values are considered to be 'high' might vary, but p–values above 0.20 are very high in most applications. We might get an almost equally good predictor by removing terms with high p-values. This can be done by editing the '`Model Template`': Place the mouse cursor in the '`Model Template`' field and click. You may now edit the work equation. Remove the term with the highest p–value and type `ctrl-RET`. Then choose '`Selected`' in the '`Analysis`' menu again. Try adding or deleting terms to find a good model. A good model is one where all terms have low p–values and Multiple R-squared (quality) is high. In other words, a good model is a model where unimportant terms are removed and the remaining terms are those that influence the response variables most markedly. Terms with high p-values might corrupt the predictor by introducing noise.

The model menu in the '`Analysis`' window also contains an option called '`Automatic Regression Model`'. Instead of manually deleting out or adding terms until you have a good model, DECISION can do this for you. If you have been editing the work equation and removed a lot of terms, choose '`Default`' in the '`Model`' menu to make the original equation appear in the '`Model Template`' field. Then change the model template to

```
AutomaticRegressionModel = (1, X0, X1, X0*X1, X0*X0, X1*X1)
```

using the '`AutomaticRegressionModel`' option of '`Model`'. The automatic regression module aims at finding the optimal model and comes out with the predictor:

```
50.5556 + 10.8333*X0 + 20.1667*X1 + 2.25*X0*X1
```

The '`Calculator`' is used to evaluate the predictor. The user provides values to the input variables listed in the lower part and enters return to obtain the corresponding values of the response variables.

Kriging is another option in the Analysis window. Kriging is a widely applied technique in Geostatistics. A kriging predictor honors the observations (as opposed to a predictor derived from regression analysis) in the sense that runs that are performed are reproduced. In other words, if X0=X1=0 is inserted into the regression predictor, the difference between the observed value and the predicted value will be the residual 50.00 - 49.44 = 0.56; the kriging predictor will produce the observed value. To run kriging on this example, use the mouse button to enter:

```
KrigingModel = (1, X0, X1, X0*X1, X0*X0, X1*X1)
```

in the '`Model Template`' field. Using the calculator you can check that the kriging predictor produces the correct value for X0=X1=0, namely 50. Consult the manual '`DECISION. Theory, Methods and Examples`' for further information and references about kriging.

The report option of the '`Analysis`' pull down menu outputs the main results of the analysis. The equation is written with standard deviation in parenthesis.

2.8 Plotting in the analysis window

Two and three dimensional plots are available. Considering two dimensional plots, the user selects the variable to be plotted on the x and y axis. A reasonable plot to make is the predicted vs. the observed values. To be specific select '`Pred`' on the x-axis and '`Obs`' on the y-axis and enter '`Ok`'. The graph window pops up. The window has a menu bar with buttons labeled '`File`', '`Edit`', '`Font`', '`Size`', '`View`'. To the right there is a column of buttons, the 4 lowest are '`Plot in 1`' to '`Plot in 4`', see Chapter 8 of '`DECISION: Reference manual`' for full details.

To view a plot, click on the '`Plot in 1`' button. A selection box appears. Click on the name of a plot (in the above case '`Pred_Obs_y1`' and '`OK`' to view it. Some further details on plots in this window follows. The names of the requested plots should be self-explanatory.

- 2_D plots.

  The analysis window contains a matrix starting with '`Id`', '`Dep.Var`'. The idea underlying the plotting options is to combine two sensible columns of the matrix to investigate visually the quality of the model or say detect outliers. It is possible to select '`Pred`', '`Obs`', '`Id`' as well as input variables to be plotted on the first axis (x-axis) and '`Pred`', '`Obs`', '`Res`' (residuals), '`St.Res`' (standardized residuals), '`Dffit`' and '`Hat`' on the second axis (y-axis), see '`DECISION. Theory, Methods and Examples`' for details.

- Response Surface (3-D plots).

  It may be of interest to view the response surface as function of the input variables. Two of the input variables are then varied whereas the user will have to fix the remaining. If there are only two input variables, as in our previous example, there is no need to assign values to them. For surfaces, the 3d plot can be rotated, panned, zoomed and stretched interactively:

  Left button
  : XY and Z rotation

  Middle button
  : Pan and zoom

  Right button
  : Y and Z stretch

When the regression analysis has been performed it is possible to simulate using the predictor derived. Open the Simulation window by clicking the '`Simulation`' button in the control panel.

2.9 The Simulation window

It is possible to simulate using only default values. Selecting 'Run Simulation' in the 'Simulation' menu suffice to perform simulation. Summary statistics appears as in the 'Simulation Results' field. The present default (standard house distribution for X0 and X1, seed 17) values and the equation above give a average value for y equal to 49.00 for the equation (1,X0,X1,X0*X1). The explanation of the remaining output ('Percentile, Std Dev, Skewness, Curtosis') may be found elsewhere, for instance in 'DECISION. Theory, Methods and Examples'.

A closer look at the input options follows:

- Number of Runs

  The number of simulations to be performed. Convenient choices are 100 or 1000; these should lead to immediate response. 10000 may lead to some waiting (less than 30 seconds). The number of simulations should be between 10 and 100000.

- Seed

  DECISION uses the seed as a starting point for generating random numbers. The seed is changed after each simulation so that new simulations will not give the same results once more. However, if you want to reproduce a previous run, you have the possibility to initialize the seed.

- Equation

  The simulation window can also be used for simulating from an equation input by the user: Write 'Eqn' in the 'Id.' field and 'X0+2*X1' in the 'Expr' field (remember to use CTRL-RET to finish) and run simulation again. A new column 'Eqn' with the simulation results from the equation appears in the Simulation Results field. This is a very useful to tool for investigating for instance the impact of assigning parameters to probability distributions. It is far more convenient to obtain summary statistics of a variable by simulation than by analytical or numerical calculation; in some cases simulation is the only practical option. The possible operators include: '+', '-', '*', '/', '^' (exponentiation), 'sin', 'cos', 'tan', 'asin' (inverse sinus), 'acos' (inverse cosinus), 'atan' (inverse tangent), 'exp' (exponential function), 'log' (natural logarithm), 'log10' (base 10 logarithm) and 'sqrt' (square root). (A complete list is included in the User's Reference manual).

- Distribution of input variables.

  Simulation results for the input variables are shown below the output for the response variables. Input variables are drawn from distributions given in the Variable Definitions window. The default distribution for the input variables is a House((-1.80902 , 0), (0,1), (1.80902, 0)) distribution (chosen to make -1 and 1 the 0.1 and 0.9 Percentiles respectively). Other distribution options include the normal distribution, discrete distribution, histogram and deterministic. Consult the manual 'DECISION. Theory, Methods and Examples' for further information about these distributions.

- Constraints on simulation In addition the user has the option to insist on the response variables being decreasing or increasing by using the 'constraints' option. The default is 'unconstrained' and should be used under normal circumstances. The default option should not be altered uncritically.
- There are two types of uncertainty that may be added to the Monte Carlo simulations: 'Model Uncertainty' and 'Statistical Uncertainty'. The former is switched on by default. The latter is off by default and abbreviated 'Stat. Uncertainty'. These options are changed by marking fields following below 'Run Simulation'. Model uncertainty adds a error term with normal distribution and standard deviation determined by the analysis while statistical uncertainty samples the coefficients from the appropriate multinormal distribution. To be specific, consider the equation

  ```
  1.0+2.1*X0
  ```

  Model uncertainty adds a normally distributed term. This term reflects the quality of the model. If statistical uncertainty is on, the equation may be considered

  ```
  a+b*X0
  ```

  where a and b have expectations 1 and 2.1 and standard deviations and correlations determined by the analysis output.

The 'Report' option in the 'Simulation' pull down menu outputs the essentials of the simulation to file.

2.10 Plots of simulation results

The simulations plots are perhaps the most useful, particularly as far as displaying results goes.

The plots are organized in a tree-structured manner, the first division being between 'Distribution', 'Density' and 'Series'. Please observe that 'Series' requires a time order to be defined in the variable window. (See Section 2.2 [Variables], page 4, for how to define the time ordering). 'Distribution' gives Cumulative Distribution Functions (CDF) and the user selects between input and response variables, see details below. The 'Density' option is treated similarly. The 'Series' option has three sub-options: 'Individual Plot', 'Box plot' and 'Percentile Plot'. More details follow.

- Distribution

  The cumulative distribution plot shows the fraction of the simulated values falling below a given y-value. These plots are named 'Cdf_'+variable name in the plot selection box.

- Density

  Density plots (histograms) of simulated values for each input or response variable. These plots are named 'Hist_'+variable name.

- Series

  The user must choose the appropriate time series.

  'Individual plot': gives simulation results for a series of response variables. If run 32 of time series 't' is chosen, the plot 't_Sim_32' is generated. Recall that these plots can only be generated if there are several response variables. See Section 3.2 [Several response variables], page 26, for an example.

  'Box plot': A boxplot shows the spread in the distribution of the simulated response variables. The middle 50% of the simulations fall within the 'box', and the vertical lines go up and down to the extremes of the data. The box is limited by the 0.25 and 0.75 Percentiles, and the median is shown as a vertical line in the middle of the box (see figure below). The Boxplot, named 'BOXPLOT' shows a boxplot of all response variables.

```
       -         Largest Observation (1.0 Percentile)
       |
       |
       |
---------------  0.75 Percentile
|             |
|             |
---------------  median            (0.5 Percentile)
|             |
|             |
|             |
---------------  0.25 Percentile
       |
       |
       |
       |
       -         Smallest observation (0 Percentile)
```

  'Percentile plot':

  Percentile plots for a series of response variables. Frequently, the user may want to plot several percentiles plot in the same figure. The appropriate plots are then selected and if time series 't' is used, enter 't_Pe*' to see all plots. Again, these plots are only generated if there are several response variables. See Section 3.2 [Several response variables], page 26, for an example.

2.11 History Matching / Optimization

This section is only relevant for readers having a history matching module. Data entered in the previous sections are not to be used in this section.

Before moving to a specific example, an outline of the contents of the History matching window is in order. The appearance of the window will of course depend on previous activity on the part of the user. However, below it is initially assumed that the user enters right into the History matching window without any prior activity. There are six pull down menus at top from left to right: '`Optimize`', '`Object Function`', '`Constraints`', '`Filters`', '`Experiments`' and '`Graphs`' in addition to '`Help`'. Following the left hand side of the window from top to bottom, there are fields called: '`History`', '`Truncation`', '`Constraints`', '`Object Function`' and '`Optima`'. At this point there is by default one response variable named '`Y0`' (with default history value 0 and weight 1) and one input variable called '`X0`' with a minimum value of -1 and maximum of 1. The functionality of the above buttons will be explained in connection with an example following shortly.

A simple example illustrating the main features in the history matching window can be loaded from the file '`prod-hist.prj`' (available from the tape).

This project file contains an example with two input variables '`x1`', '`x2`', and 6 response variables '`y1`', '`y2`', '`y3`', '`y4`', '`y5`', '`y6`'. By opening the design window the user can examine the design. Analysis has been performed, and the user can check the results from analysis by opening the analysis window. Now let's turn to the history matching window:

In the history field history values have been entered for the response variables y1 – y6:

```
           @y1   @y2   @y3   @y4   @y5   @y6
History:   17.4  21.8  25.1  27.7  29    30.8
```

In the object function field, a list of object functions is visible. The first one reads

```
F1    (1*(@y1-y1)^2+1*(@y2-y2)^2+1*(@y3-y3)^2+1*(@y4-y4)^2+1*(@y5-y5)^2+1*(@y6-y6)^2
```

Select this object function by clicking on it with the left button. Then select '`Optimize All`' in the '`Optimize`' menu. The results are shown in the '`Optima`' field:

```
Id  Value     x1        x2        y1        y2        ...
 1  0.003483  0.513082  -0.823109 17.372100 21.791800 ...
 2  0.009499  0.543446  -0.853647 17.368400 21.806800 ...
 3  0.209885  0.114469  -0.391572 17.376100 21.626500 ...
 .  .         .         .         .         .         .
 .  .         .         .         .         .         .
 .  .         .         .         .         .         .
```

(you may need to scroll to see all the numbers).

'Id' indicates the ranking of the matches with 1 being the most promising candidate. 'Value', the value of the objective function, gives a measure of quality for the match, the lower the better. Next, the suggested values for the input variables follow.

A few words on the method: The object function may have several local optima, therefore a set of different starting points are used. The set of starting points is obtained by evaluating the object function on a grid and keeping those that give the smallest value for the object function. Numerical optimization from these starting points is done using a conjugate directions method. To constrain optimization to the region defined by the intervals given for each input variable, one may add a "penalty" function to the object function. The penalty adds a quadratic term to the object function if x is outside the region.

In this example, a penalty function named 'P' has been defined, and if you scroll the object function field, you will also find the object function 'F2' which is the previous object function 'F1' plus 'P'. In this example, 'P' constrains optimization to the intervals [-1,1] for both input variables. To perform constrained optimization, click on the line that reads

```
F2        F1 + P
```

and select 'Optimize all' in the 'Optimize' menu. The results should be the same as above. (The results would differ if the response surface had an optimum outside the region defined).

2.11.1 Plotting

To plot the history, select 'History' in the 'Graph' menu. Select the series named 'y' in the selection box that appears, and click on 'OK'. The Graph window will now appear if it is not already present (i.e. has been activated by your earlier activity). Click on 'Plot in 1' in the Graph window, select 'History_y' and click on 'OK' to view the plot. To compare this to the predicted values in the optimum, select the first optimum in the 'Optima' list by clicking on it (the line will be highlighted). Then select 'Optimum' in the 'Graph' menu, select the series named 'y' in the selection box that appears, and click on 'OK'. To plot history and optimum together, click on 'Plot in 1' in the Graph window and write 'History_y Opt_1_y'. As you can see from the plot, this is a good match (but remember, this is a synthetic example, real-world situations are often more complicated).

It is also possible to plot the object function surface. Select an object function (e.g. 'F2') by clicking on it, then choose 'Surface' in the 'Graphs' pull down menu.

2.11.2 Additional runs

It will often be necessary to perform additional runs to improve the quality of the match. This will be an iterative process consisting of the following steps.

1. Add new experiments to the design.
2. Run the new experiments.
3. Rerun analysis.
4. Perform optimization.

If DECISION is combined with an external program by the use of filters, in such a way that it is possible to start the external program from DECISION (see 'DECISION: Reference Manual' for how to do this, the setup is probably a task for a local expert), the loop above can be performed automatically by selecting 'Set Filter' in the 'Filter' menu. The user can then set the number of new experiments to be added to the design in each iteration, and the number of iterations of the loop.

Iterations can also be performed manually; then the procedure will be as follows:

1. Open the Variables window. If you want different levels for the new experiments, change levels.
2. Open the Design window. Generate Candidates. Increase the size of the design by increasing the number following 'Dopt(' in the selection field. Then click on 'Select (All)' in the 'Design' menu.
3. Run the new experiments
4. Open the analysis window and perform analysis.
5. Open the history matching window again and perform optimization as described above.

2.11.3 Report

The 'Report' option of the 'Optimize' pull down menu is used to get output of the history matching window to file.

2.11.4 Experiments

It is possible to write the save, run and load the experiments in the list of optima by use of the corresponding functions in the 'Experiments' pull down menu. These functions have the same functionality as in the design window, see Section 2.5 [Candidate set], page 7. Also, if you save, run or load an experiment from the History Matching window, it will automatically be added to the selected design in the Design window.

2.12 Further Graph options

For convenience, we reproduce some details from 'DECISION: Reference manual' here. Note that the first time plotting is activated, the window appears. Subsequent plots are sent without delay and the window appears until it is quitted from the graph window. If it is desired to make several plots in one larger, using say 'Plot in 4', it is preferable to first select all plots and not change between the graph window and say simulation window. *Several plots together:* Type several names in the text field in the plot selection box (ex: 'Perc_10 Perc_50 Perc_90'). Wildcard ('*') can also be used (ex: 'Perc_*'). A useful plot is obtained by 'Cdf_YO Hist_YO').

- Filter

  The plot selection box might contain a large number of plot names. Use a filter to select a certain group of plots (ex: Type in 'C*' and click on 'Filter' to get all plots starting with a 'C'. Type '*' and click on 'Filter' to get back to the list with all plots.)
- Split.

  Split screen in 2-4 plotting windows: Click on 'Plot in 2' - 'Plot in 4'.
- Print plot

  Click on 'File' in the upper left corner and select 'Print' in the menu that appears. This will generate a PostScript file of the plot. The PostScript files will be named 'plot001.ps', 'plot002.ps' and so on.
- Add/remove text from plot

  To remove text, click on 'Edit' in the upper left corner and select 'Cut' in the menu that appears. Then click on the left end of the text.

To add text, click on ‘Edit’ and select ‘Paste’ in the menu that appears. Type in your text in the box that appears (default is the text that was last cut) and point and click with the mouse to place it.

- Clear Screen

  Click on ‘View’ in the upper left corner and select ‘Clear’ in the menu that appears.

- Set axes.

  To set your own axes in a plot, turn off the default before selecting the plot. Turning off the default axes is done by clicking on the ‘Default on’ button to the left in the window. The text on the button will be ‘Default off’ when it has been turned off. When you now select a plot using for instance ‘Plot in 1’, you will be asked to set the axes. Three numbers should be entered, min, max, and a number that controls where tick-marks should be placed.

- Delete.

  The ‘Delete’ button deletes a plot from the plot list.

- Change type.

  The ‘Change type’ may be used to display plots differently. To remove the lines interpolating a cdf plot, enter ‘Change type’ followed by the plot to change. A number of options then appear. In this case choose scatter. Replot to see the only the points.

2.13 Files

If you want to save the results of your work so far, you can save your project in a file. Saving and loading files is done in the ‘Project Menu’ in the control panel.

If this is a new project, it must be given a filename before saving. Select ‘Save As’ in the ‘Project’ menu, select or enter a filename in the file selection box that pops up, and click the ‘OK’ button.

To load a file, select ‘Load’ in the ‘Project’ menu. A file selection box pops up. In the right part of this window is a list of files. You can select a file by clicking on it. The filename will be highlighted. Then click on the ‘OK’ button in the lower left corner of the window. DECISION will read the file, and the file selection box will disappear.

2.14 Ending your DECISION session

If you want to continue working on your project in DECISION on a later occasion, remember to save it before leaving DECISION.

To leave DECISION, select '`Quit`' in the '`Project`' menu on the control panel.

2.15 Summary of DECISION windows

Here is a short summary of the main activities in the various windows of DECISION:

Project window

- Load/save project files.
- Open other windows.

Variables window

- Define variables.
- Define levels that will be used in design.
- Define distributions that will be used in Monte Carlo simulation.

Design window

- Generate Candidate set.
- Generate Design (a selection from the candidate set).
- Enter Response values

Analysis window

- Generate Regression, Automatic Regression or Kriging models.
- Automatic Regression can be used for model selection (i.e. selection of terms to be included in regression model).
- The equation generated is used in the simulation window and/or the history matching window.

Simulation window

- Generates means, standard deviations and percentiles of response variables by Monte Carlo simulation.
- Results can be presented by various plots.

History Matching window

- Uses equation(s) generated in the Analysis Window
- Optimization using various user specified object functions

Plotting

- The Graph window is activated from the other windows by selecting a plot.

3 Examples

3.1 SPE example

This section presents a realistic example based on Damsleth, Hage and Volden (Offshore Europe, SPE 23139, 1991). In addition to demonstrating a case study, we show some of the more advanced features of DECISION. A very brief description of the case study and the variables follows (full details are provided in the mentioned paper):

- PR Production after four years (response).
- ST STOOIP.
- LO Lobes.
- RP Relative permeability.
- SK Skin.
- KV Vertical permeability.

Values of PROD are ST scaled to maintain confidentiality. All variables are coded; -1 (+1) corresponds to the low (high) value while the base case is coded as 0. We concentrate on analysis and simulation. The data is loaded from the file '`spe.prj`'.

Analysis

First analyze the data using the `Analysis` module. Some of the vital output is

```
20.8715*1 + 4.79114*ST + 0.465125*LO + 0.92726*RP + 0.475748*SK
                                                       + 0.707668*KV
*Warning*  1 experiment is outlier (number 5)

Residual Standard Error = 3.21187

  Term    coef std.err  p.value
     1   20.87    0.92    0.00
    ST    4.79    1.16    0.00
    LO    0.47    1.09    0.68
    RP    0.93    1.15    0.44
    SK    0.48    1.40    0.74
    KV    0.71    1.64    0.68
```

This output and the "quality" of the model (R-square, which is equal to 0.763673) indicate that it is worthwhile trying to find a better predictor; the Multiple R-square could be increased. Several of the input variables have high p–values indicating that they could be excluded. Note however that exclusion of variables or terms will not increase the R-square value. Based on the p–values, the intercept and ST should be retained. In addition one could try an interaction term. Editing the equation ending up with

```
RegressionModel = ( 1, ST, ST*LO )
```

and re running the regression, the predictor becomes

```
20.9044 + 4.569*ST + 1.97707*ST*LO
```

All terms have p–values below 0.05 and the multiple R-square has been increased to 80.9 %. This is a reasonable model. The final model in Damsleth et al. includes several interaction terms and a nonlinear term:

```
RegressionModel = ( 1, ST, LO, RP, ST*LO, LO*RP, (ST*ST+LO*LO+RP*RP) )
```

The predictor based on this equation is

```
23.9901 + 4.7094*ST + 0.715683*LO + 0.967896*RP +
1.9179*ST*LO - 1.0571*LO*RP - 1.34871*(ST*ST+LO*LO+RP*RP)
```

This equation will be used for the coming Simulation (slightly differing equations obtained by using the automatic regression option may well be equally good). The user is warned that observation no. 9 is an outlier. At this point the model may be investigated using the plotting facilities provided by DECISION.

Simulation

Our '`surrogate simulator`' involves the three variables ST, LO and RP. These variables have been assigned distributions coinciding with the ones in Damsleth et al.: ST and RP are house distributed on [-1,1], the chance of the endpoints should be 0 while the chance of 0 (the middle value) could be any positive number (scaling is performed automatically). Finally, LO is discrete with probabilities 0.2,0.6,0.2 for the values -1,0 and 1 respectively. After having provided the distributions and run the simulation, we find that (the numbers in parenthesis are from Damsleth et al.) the mean is 22.95 (22.99) and the standard deviation 2.78 (2.80). The 10000 simulations are confined to the interval [13.22, 32.26] ([12.74,31.12]). Different simulations produce different numbers; to ensure equality the random seed must be initialized to the same number. In our simulation above the seed value was set to 17. (The simulated values may unfortunately depend on the distribution of input variables not used in the predictor. This will be changed in future releases.)

3.2 Several response variables; uncertainty in a production profile

In many applications it is natural to have several response variables all depending on the same set of input variables. In the previous example we might not only be interested in production after 4 years, but also after 5,6,7,... years.

A simple example with several response variables can be loaded from the file '`prod_profile.prj`' (available from the tape). This file contains a project where all variables have been defined, the design has been selected and response values have been entered. To calculate predictors for all response variables simultaneously, select '`Analyze all`' in the Analysis menu. You may now examine the results for each response variable by clicking on the corresponding line in the '`Response Variables`' field in the upper part of the window.

The simulation output from 1000 simulations with seed 17 and X1, X2 both house distributed with parameters ((-1,0), (0,2), (1,1), '`Model uncertainty`' on, is:

Response Variables:

Perc	y1	y2	y2	y4	y5	y6
0.00	13.77	18.51	19.67	21.82	24.02	23.94
0.10	16.13	20.39	22.57	25.49	26.81	28.49
0.20	16.58	20.88	23.35	26.38	27.84	29.59
0.25	16.76	21.04	23.62	26.71	28.16	29.97
0.50	17.38	21.87	24.93	28.05	29.82	31.93
0.75	17.97	22.79	26.23	29.34	31.72	33.90
0.80	18.14	22.96	26.50	29.66	32.10	34.30
0.90	18.51	23.49	27.26	30.30	33.06	35.45
1.00	19.68	25.17	30.56	32.43	36.23	39.40
Average	17.33	21.90	24.93	27.96	29.90	31.91
Std.Dev.	0.95	1.19	1.82	1.88	2.41	2.71

The simulation results can preferably be examined by plotting as explained previously. See Section 2.10 [Simulation Plotting], page 16.

4 How to select probability distributions

DECISION offers a large variety of input variable distributions. This chapter is intended to give inexperienced users some hints on how to select probability distributions. The first section consists of different questions that are intended to help inexperienced users choose distribution. The second section lists and describes the available distributions, while the third section documents how dependencies between variables can be defined in DECISION. Finally, we have a section on terminology that tries to give a short explanation of some frequently used words in statistics.

4.1 A Guide to the distributions.

Random or deterministic?

Is the variable random or deterministic? (Deterministic = fixed, constant). See Section 4.2.13 [Deterministic Distribution], page 36 for the deterministic option, if not, go on reading:

Discrete or continuous?

If the variable is random, the next question is whether it has a discrete or continuous distribution?

Discrete: It has a discrete distribution if it can only take a finite number of values. An example of a discrete distribution is throwing dice; you only get the discrete values 1, 2, 3, ..., but never 2.4. On the other hand, a variable that has a continuous distribution can take all values in an interval. See Section 4.2.12 [Discrete Distribution], page 36 for how to specify a discrete distribution.

Continuous:

A continuous distribution can take all values in an interval. Continuous distributions can be divided into two groups: Symmetric and skewed (asymmetric) distributions. See Section 4.1.1 [Symmetric Distributions], page 29 for symmetric distributions. See Section 4.1.2 [Skewed Distributions], page 30 for skewed distributions.

```
|         X                  |
|    X    X    X             |   xXXXx.
|    X    X    X             |  XXXXXXXXXx.
|    X    X    X             | XXXXXXXXXXXXXX
|------------------>         |------------------>
      Discrete                     Continuous
```

The following figure gives a quick overview:

```
                Distribution
                     /\
                    /  \
                   /    \
                  /      \
                 /        \
                /          \
               /            \
          Discrete       Continuous
                             /\
                            /  \
                           /    \
                          /      \
                         /        \
                        /          \
                       /            \
                      /              \
                     /                \
                 Symmetric          Skewed (asymmetric)
                   /                  /\
                  /                  /  \
                 /                  /    \
                /                  /      \
               /                  /        \
              /                tail        tail
             /              towards        towards
            /\             the left        the right
           /  \               /\              /\
          /    \             /  \            /  \
         /      \           /    \          /    \
        /        \         /      \        /      \
   Limited      not       (L)     (N)     (L)    (N)
    to an     limited      |               |      |
   interval     (N)        |               |      |
     (L)         |         |               |      |
      |          |         |               |   Lognormal
      |          |         |               |       or
    Trunc.     Normal      |               |     Gamma
    Normal                 |               |
      or                 House           Trunc.
    Uniform               or           Lognormal
      or                 Beta             or
    House                                House
      or                                  or
     Beta                                Beta
```

If none of these fit, any distribution can be specified using a Histogram distribution.

4.1.1 Symmetric Distributions

Belonging to this group are the Normal distribution, the truncated Normal distribution, the Uniform distribution, symmetric House distributions and symmetric Beta distributions.

Limited to an interval, or not?

That is, are there any sharp limits that the variable can not fall outside?

Not limited to an interval:

The Normal distribution is bell-shaped, characterized by its mean and standard deviation. Even if 95 percent of the observations will fall within +/- twice the standard deviation, it is not limited to an interval. See Section 4.2.1 [Normal Distribution], page 31.

Limited to an interval:

The rest of the symmetric distributions mentioned, are limited to an interval.

Bell-shaped distribution, limited to an interval:

Try the truncated Normal distribution, with the same shape as the Normal distribution, except that upper and lower limits are specified. See Section 4.2.2 [Truncated Normal], page 31.

All values equiprobable?

If all values are equiprobable, choose the Uniform distribution. See Section 4.2.11 [Uniform Distribution], page 35.

Triangle-shaped?

If you want to specify a triangle-shaped distribution, choose the House distribution, and set parameters so that it is symmetric. See Section 4.2.9 [House Distribution], page 34.

Limited to values between 0 and 1:

If the distribution is limited to the interval [0, 1], Beta distributions can also be used. A wide variety of shapes can be obtained with different choices of the parameters. To obtain a symmetric distribution, the two parameters of the Beta distribution should be identical. See Section 4.2.5 [Beta Distribution], page 33.

If nothing else fits:

Finally, if none of the above distributions fit, you can always specify any distribution with the Histogram distribution, see Section 4.2.14 [Histogram Distribution], page 37.

4.1.2 Skewed Distributions

Skewed distributions are the asymmetric distribution alternative. Belonging to this group are Lognormal and truncated Lognormal distributions, Gamma distributions, and asymmetric House distributions and Beta distributions.

If the distribution has a tail extending towards the right (towards more positive values) we say that it is right (positive) skewed. And if the distribution has a tail extending towards the left (towards more negative values) we say that it is left (negative) skewed.

```
|                        |                        |
|   xXXx.                |      xXXXXx            |             .xXXx
|  XXXXXXXXx.            |   xXXXXXXXXXXx         |         .xXXXXXXXX
| XXXXXXXXXXXXXXX        |  XXXXXXXXXXXXXX        |  XXXXXXXXXXXXXXX
|------------------>     |------------------>     |------------------>
  Positive (right)             Symmetric            Negative (left)
    skewed                                             skewed
```

Distributions with left (negative) skewness:

If it is skewed towards the left, try a skewed house distribution (see Section 4.2.9 [House Distribution], page 34). Or if the variable is limited to [0, 1], a Beta(`a`,`b`) distribution where `a>b` is a good alternative. See Section 4.2.5 [Beta Distribution], page 33.

Distributions with right (positive) skewness:

Limited downwards but not upwards:

If it is skewed towards the right and limited downwards but not upwards, try a Gamma distribution or a Lognormal distribution. See Section 4.2.6 [Gamma Distribution], page 33 for how to specify the Gamma distribution. See Section 4.2.3 [Lognormal Distribution], page 32 for how to specify the lognormal distribution.

Limited upwards and downwards:

If it is skewed towards the right and limited both upwards and downwards, use a truncated lognormal distribution or a house distribution. If it is limited to the interval [0, 1], you may also use a Beta(`a`,`b`) distribution with `a<b`. See Section 4.2.4 [Truncated Lognormal Distribution], page 33 for how to specify the truncated lognormal distribution. See Section 4.2.9 [House Distribution], page 34 for house distribution and see Section 4.2.5 [Beta Distribution], page 33 for the Beta Distribution.

If nothing else fits ...

If none of the above distributions fit, you can always specify any distribution with the Histogram distribution, see Section 4.2.14 [Histogram Distribution], page 37.

4.2 Available Distributions

This list gives a short description of available distributions in DECISION. For formulae, (e.g mean and variance) not contained here, consult the manual '`DECISION: Theory, Methods and Examples`' or a statistics textbook.

4.2.1 Normal Distribution

In DECISION, the Normal distribution is specified with:

```
Normal(mean, std. deviation).
```

The distribution is symmetric and bell-shaped. In the long run, approximately 95% of the observations should lie within the interval given by the mean plus/minus twice the standard deviation. The Normal distribution is also sometimes called the Gaussian distribution.

```
|
|
|                         xXXx
|                       xXXXXXXx
|                     xXXXXXXXXXXx
|                    xXXXXXXXXXXXXx
|                   xXXXXXXXXXXXXXXx
|                 xXXXXXXXXXXXXXXXXXXx
|               xXXXXXXXXXXXXXXXXXXXXXXx
|      ....xxXXXXXXXXXXXXXXXXXXXXXXXXXXXXXxx....
|------------------------------------------------------------>
                 The Normal distribution
```

4.2.2 Truncated Normal Distribution

The user can modify the normal distribution by truncating values

```
TruncatedNormal(mean, std. deviation, lower limit, upper limit)
```

The parameters '`mean`' and '`std. deviation`' are the mean and standard deviation of the untruncated distribution, the actual mean and standard deviation are changed by the truncation.

```
|                        xXXx
|                      xXXXXXXx
|                    xXXXXXXXXXXx
|                   xXXXXXXXXXXXXx
|                  xXXXXXXXXXXXXXXx
|                xXXXXXXXXXXXXXXXXXX
|                XXXXXXXXXXXXXXXXXXX
|                XXXXXXXXXXXXXXXXXXX
|----------------|-----------------|------------------------>
                 t2                t1
         The Truncated Normal distribution
```

```
TruncatedNormal(mean, std. deviation, t1 , t2)
```

4.2.3 Lognormal Distribution

A variable has a lognormal distribution if the logarithm of the variable is normally distributed. The lognormal distribution is specified by:

```
Lognormal(mean log, std. deviation log)
```

where '`mean log`' and '`std. deviation log`' are the mean and the standard deviation of the *logarithm* of the variable. The lognormal distribution takes only positive values and is skewed to the right.

```
|     XXx
|    XXXXXx
|   .XXXXXXX
|   XXXXXXXXXx.
|  .XXXXXXXXXXXXx.
|  XXXXXXXXXXXXXXXXXxx..
| .XXXXXXXXXXXXXXXXXXXXXXXxx...
|.XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXxxx.......
|XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXxxxxx
|------------------------------------------------------------->
   The Lognormal distribution
```

4.2.4 Truncated Lognormal Distribution

The lognormal distribution can also be truncated:

```
Truncated Lognormal(mean log., std. deviation log., lower limit, upper limit)
```

where '`mean log`' and '`std. deviation log`' are the mean and the standard deviation of the *logarithm* of the variable, and '`lower limit`' and '`upper limit`' are upper and lower limits of the logarithm of the variable. Note that a lognormally distributed variable takes only positive values, even if there is no truncation.

4.2.5 Beta Distribution

```
Beta(shape parameter a, shape parameter b)
```

The beta distribution takes values on the interval [0,1]. If `a=b`, the distribution is symmetric. If `a<b`, the distribution is skewed towards the right, and if `a>b`, the distribution is skewed towards the left. '`DECISION: Theory, Methods and Examples`' contains plots showing examples of beta distributions with different choices of the parameters and gives formulae for mean and variance, and for estimating the parameters a and b if mean and variance are known.

```
|          a<b          |              a=b       | a>b
|  xXXx.                |    xXXXXXx             |        .xXXx
| XXXXXXXXXx.           | xXXXXXXXXXXXx          |    .xXXXXXXXX
|XXXXXXXXXXXXXXXX       |XXXXXXXXXXXXXXXX        |XXXXXXXXXXXXXXXX
|--------------|-->     |--------------|--->    |--------------|--->
0              1        0              1        0              1


|X          b>a         |X    a=b=0.5    X       |     a=b=1
|XXx.       a=1         |X.            .X       |
|XXXXXXx.               |XXx.        .xXX       |XXXXXXXXXXXXXXXX
|XXXXXXXXXXXXxxxx       |XXXXXxxxxxXXXXX        |XXXXXXXXXXXXXXXX
|--------------|-->     |--------------|--->    |--------------|--->
0              1        0              1        0              1
```

Beta distribution examples

4.2.6 Gamma Distribution

```
Gamma(shape parameter t, scale parameter p, left endpoint x0)
```

The shape parameter `t` determines the shape of the distribution. If `t < 1`, the probability density tends to zero as x tends to the left endpoint `x0`. If `t > 1`, the distribution has a bump at `x0 + (t-1)/p`. The scale parameter `p` stretches the distribution, and `x0` is the left endpoint of the distribution (i.e, no observations below `x0`). '`DECISION: Theory, Methods and Examples`' contains plots showing examples of gamma distributions with different choices of the parameters and gives formulae for mean and variance, and for estimation of parameters `t` and `p` if mean and variance are known.

```
|X       t<1                |   xx      t>1
|Xx                         |  XXXXx.
|XXXx.                      | xXXXXXXXXx.
|XXXXXXxxx...               |.XXXXXXXXXXXXxx...
|------------------>        |------------------>
        Gamma distribution examples
```

4.2.7 Chi-Square Distribution

The Chi-square distribution with `v` degrees of freedom is a Gamma(`v`/2, `v`/2, 0) distribution.

4.2.8 Exponential Distribution

The Exponential distribution is a Gamma(1,p,`x0`) distribution.

4.2.9 House Distribution

```
House( (x1, y1), (x2, y2), (x3, y3))
```

The parameter `x1` is the minimum x-value, `x2` is the x-value of the mode and `x3` is the maximum x-value, and `y1`, `y2` and `y3` are the corresponding y-values.

```
|                               y2
|                               .
|                            .xXXXx.
|                         .xXXXXXXXXXx.
|                      .xXXXXXXXXXXXXXXXx.
|                   .xXXXXXXXXXXXXXXXXXXXXXx.
|               y1 XXXXXXXXXXXXXXXXXXXXXXXXXXX y3
|                  XXXXXXXXXXXXXXXXXXXXXXXXXXX
|                  XXXXXXXXXXXXXXXXXXXXXXXXXXX
|------------------|------------|------------|-------------->
                   x1           x2           x3

         A House((x1,y1),(x2,y2),(x3,y3)) distribution
```

4.2.10 Triangle Distribution

The triangle distribution is a special case of the house distribution:

```
House( (x1, 0), (x2, y2), (x3, 0))
```

The manual '`DECISION: Theory, Methods and Examples`' contains formulae for calculating `x1` and `x3` if `x2` and two percentiles on each side of `x2` are known.

```
|                               y2
|                               .
|                            .xXXXx.
|                         .xXXXXXXXXXx.
|                      .xXXXXXXXXXXXXXXXx.
|                   .xXXXXXXXXXXXXXXXXXXXXXx.
|-------------------|-----------|-----------|-------------->
                    x1          x2          x3
                      Triangle distribution
                   House((x1,0),(x2,y2),(x3,0))
```

4.2.11 Uniform Distribution

A Uniform distribution on [`x1`, `x3`] can be defined by

```
House( (x1, 1), (x2, 1), (x3, 1))
```

where `x1` is the left endpoint of the distribution, `x3` is the right endpoint, and `x2` is any number such that `x1 < x2 < x3`.

```
|                  XXXXXXXXXXXXXXXXXXXXXXXXXXX
|                  XXXXXXXXXXXXXXXXXXXXXXXXXXX
|                  XXXXXXXXXXXXXXXXXXXXXXXXXXX
|------------------|------------|------------|-------------->
                   x1           x2           x3
                      Uniform distribution
                   House((x1,1),(x2,1),(x3,1))
```

4.2.12 Discrete Distribution

A variable has a discrete distribution if it only takes the discrete values `x1`, `x2`, ..., `xN`, with 'chances' `p1`, `p2`, ..., `xN`. In DECISION, the discrete distribution is specified with:

```
Discrete((x1, p1), (x2, p2), ... , (xN, pN))
```

We use 'chance' rather than probability since the numbers are un-scaled. The scaling is done automatically by DECISION

```
|                      p2 X
|                         X            p3 X
|              p1 X       X               X
|                 X       X               X
|                 X       X               X
|                 X       X               X
|-------------------|--------|------------|--------------->
                   x1       x2            x3
      Discrete((x1, p1), (x2, p2), (x3, p3)) distribution
```

4.2.13 Deterministic Distribution

The deterministic distribution is specified with:

```
Deterministic(value)
```

The deterministic distribution can be used to fix a variable to a constant value.

4.2.14 Histogram Distribution

A histogram distribution can be specified from a table of percentiles:

```
xmin   0
x1     y1
x2     y2
 :     :
 :     :
xmax   1
```

where y1 is the cumulative probability that an observation is less than x1 (in other words, x1 is the y1-percentile), y2 is the cumulative probability that an observation is less than x2, and so on. xmin is the 0 percentile (left endpoint of the distribution) and xmax is the 1 percentile (right endpoint). The histogram distribution can be entered into DECISION as follows:

```
Histogram((xmin, 0), (x1, y1), (x2, y2) ... , (xmax, ymax))
```

Note that the parameters must satisfy `xmin < x1 < x2 < ... < xmax` and `0 < y1 < y2 < ... < ymax`. If `ymax` is not equal to 1, the 'chances' `y1`, `y2`, ... will be scaled automatically by DECISION.

4.3 Dependent input variables

Using dependent variables is a slightly more advanced use of DECISION and this section therefore mostly intended for the experienced users.

Dependent probability distributions can be specified by letting a parameter to a distribution depend on other input variables in a sequential manner. The parameters to the distributions need no longer be numbers, but they can be any expression, also expressions containing other input variables previously defined.

For example, it is possible to specify

```
Name  . . . . . .  Distribution
X0    . . . . . .  Normal(0, 1)
X1    . . . . . .  Normal(X0, 1)
```

or one can introduce a dummy variable

```
Name  . . . . . .  Distribution
U     . . . . . .  Normal(0, 1)
X0    . . . . . .  Normal(2, 1) + 0.5^(1/2) * U
X1    . . . . . .  Normal(3, 2) + 0.5^(1/2) * U
```

to achieve covariance 0.5 between `X1` and `X2`.

Note that using expressions instead of numbers as parameters to the distributions may lead to inconsistent definitions if it is not used correctly. Things that won't work include for example specifying an expression for the variance that will sometimes be negative since negative variances are not defined.

4.4 Terminology

Chance Used in DECISION for un-scaled probability (or probability density).

Mean Average or expected value. Sometimes there is a slight difference in meaning between '`mean`' and '`expected value`', where '`mean`' is the value calculated from data, while '`expected value`' is the theoretical value.

Median The value that 50 percent of observations fall below; the 0.50 percentile.

Mode A maximum of the probability density.

Percentile A number X is the t-percentile of a distribution if the t percent of the observations from this distribution are smaller than X.

Standard deviation
Standard error (used interchangeably).

Truncation
Rejection of values outside specified limits.

Variance Squared standard deviation.

Concept Index

P

Q

R

S

T

U

V