Extension

The SHELF 3 Extension template provides a flexible tool to address a variety of elicitation situations. The principal features of extension are as follows.

1. Two uncertain quantities. One is referred to as the *target* quantity, and will be denoted here by X. The other is the *extension* variable, denoted here by Y.
2. Two elicitations are carried out. One is to elicit the (marginal) distribution of Y. The other is to elicit the (conditional) distribution of X given particular values of Y.
3. The purpose of the extension may be to elicit the distribution for X, in which case extension is an elaboration (known as extending the argument to include Y) that is useful when it is difficult to elicit the distribution of X directly. The purpose may alternatively be to elicit the joint distribution of X and Y when they are not independent, in which case extension is one of several multivariate elicitation methods in SHELF.

We will use the following examples to illustrate the various uses of the Extension template.

Example 1. A company is considering launching a new product line. An elicitation workshop is to be convened to provide expert judgements about the value of sales of the new product in each of the first two years after the launch. These two quantities are not independent; the experts feel that it would be easier for them to judge the value of sales in the second year if they knew the value in the first year. Therefore, we define the target quantity X to be the value of sales (in millions of US dollars) in the second year and the extension variable Y to be the value in the first year. The objective in this example is multivariate elicitation, i.e. to elicit a joint distribution for X and Y.

Example 2. The quantity of interest in this example is the time X (in days) taken to complete a building project. In considering their judgements about X, it would be easier if the experts knew the daily temperatures, because work will proceed more slowly in cold weather, and even more slowly if the ground is frozen. It is decided to use the extension elaboration with the extension variable Y being the average daily temperature during the building project. A separate workshop will be convened using experts in local weather forecasting to elicit a distribution for Y.

Example 3. An expert panel is convened to provide judgements about the rainfall X in the coming year in South Africa. The experts feel that it could be high, 800 millimetres or more, or a more moderate 300-500mm because rainfall in South Africa is strongly influenced by the El

Niño Southern Oscillation (ENSO). The country receives lower than average rainfall in El Niño years and higher than average in La Niña years. This suggests using the Extension template with extension variable Y, where Y can take three possible values – El Niño, neutral and La Niña.

Example 4. A doctor suspects that a patient may have kidney disease and sends a urine sample to the pathology laboratory for testing. Letting X be the result of the test, the distribution of X will clearly depend on whether the patient has kidney disease or does not, so we let Y be the disease state with those two possible values. However, in this example the ultimate objective is to assess a probability for the patient having kidney disease conditional on the outcome of the urine test.

The way that the Extension template is used depends very much on the nature of X and Y, in particular on whether they are continuous or discrete. Extension can even be used when either X or Y or both are multivariate, i.e. made up of two or more quantities, but we begin by assuming that they are both univariate, i.e. single quantities, as in the above examples.

Case A: Continuous *X*, continuous *Y*

In most situations, expert elicitation is required for continuous quantities of interest, so we will first consider the case where both X and Y are continuous. Guidance notes on each of the principal fields of the template are presented in order below.

The guidance is illustrated using Example 2 (time to complete a building project). In addition, the Samples folder in the SHELF package contains a completed SHELF 3 Extension template for Example 1 (sales of a new product)

All elicited judgements are those of RIO, the Rational Impartial Observer, and are obtained as usual after experts make individual expert judgements followed by discussion. Unless specified otherwise, the individual judgements, discussion and RIO consensus judgements should all be recorded.

The "Extension variable distribution" field

Here should be recorded the elicited distribution for Y. Since Y is continuous, this will generally mean completing a SHELF 2 template for Y. (Note, however, that extension may also be useful for this, and so a SHELF 3 Extension template might be used to elicit the distribution of Y by extending the argument to include another extension variable. See the discussion of "Chained quantities" later in this document.) Just the distribution is recorded in this field, with the SHELF 2 record of the elicitation being attached.

The "Conditioning points" field

In principle, we need an elicited conditional distribution of X for every possible value of Y, but this is not feasible when Y is continuous and can thereby take any possible value in some range. We therefore elicit judgements about X conditional on only a selection of the possible values of Y. These are called the conditioning points. It is important for values to be chosen so that there is enough difference between them for the experts to feel able to make meaningful judgements about how the distribution of X changes as the conditioning point changes.

We recommend using five conditioning points – the median, the quartiles, and the 5th and 95th percentiles of the elicited distribution of Y. We denote these points in increasing order by $Y1$, $Y2$, $Y3$, $Y4$ and $Y5$. Thus, $Y1$ is the 5th percentile, $Y2$ the lower quartile, $Y3$ the median, $Y4$ the upper quartile and $Y5$ the 95th percentile.

The "Target quantity distributions" field

Here, conditional distributions of X given individual values of Y are elicited and recorded. It is important that the experts understand that when making conditional judgements they are to assume that Y takes the value of the relevant conditioning point and to make judgements conditional on that assumption.

This is the most complex part of the Extension method, particularly in this Case A, when both X and Y are continuous. We first present a basic recommended approach, but this is followed by a discussion of variations that might be considered.

Basic approach. The following six-step procedure is recommended for when both X and Y are continuous. Following the recommended choice of conditioning points, we suppose that there are five points denoted $Y1$ to $Y5$.

Outer medians: Elicit and record median values of X conditional on $Y = Y1$ and $Y = Y5$. These values, $m1$ and $m5$ respectively, characterise the range of influence of Y on the experts' opinions about X.

$Y3$-distribution: Elicit and record the distribution of X given that $Y = Y3$. We will refer to this distribution as the $Y3$-distribution. Just the distribution is recorded here, with the SHELF 2 record of the elicitation being attached. However, the median of the distribution, $m3$, should also be recorded here.

It is important for the facilitator to explain clearly to the experts the distinction between (a) the range of median values of X as Y varies, as characterised by $m1$ and $m5$, and (b) the range of plausible values of X conditional on Y being fixed at $Y3$. A PowerPoint template "Conditional range" is provided in the SHELF package to assist with this task.

Inner medians: Elicit median values $m2$ and $m4$ for X conditional on $Y = Y2$ and $Y = Y4$, respectively.

Transformation: The last two steps of the basic approach will yield different results if a transformation is applied to X. There are two standard transformations.

- If X is constrained to lie above some physical or logical lower bound, a log transformation is indicated. The most usual lower bound is 0, i.e. X can only take positive values, and the transformation is simply to apply the final steps on the scale of $X^* = \log(X)$. (If the lower bound is b then we define $X^* = \log(X - b)$.)
- If X is constrained to lie between some physical/logical upper and lower bounds, a logistic transformation is indicated. The most usual bounds are 0 and 1, e.g. X is a proportion (or 0 and 100 if X is a percentage). The transformation then is $X^* = \log\{X/(1 - X)\}$. (If the bounds are b and c, then $X^* = \log\{(X - b)/(c - X)\}$.)

When X has no physical or logical bounds then no transformation is indicated. Otherwise, it is recommended always to use the indicated transformation, even though a log or logistic transformation may sometimes yield almost identical results to no transformation.

The SHELF software automatically takes account of log or logistic transformations, so it is not necessary to compute values of X^*.

Median model: The next step begins the process of using the elicited judgements at the conditioning points to form the conditional distribution of X conditional on all possible values y of Y, by defining a function $m(y)$ to represent the median of X conditional on $Y = y$. In the basic approach, we simply interpolate the elicited medians $m1$ to $m5$ by a piecewise-linear function, as shown in the example below.

(If a log or logistic transformation is used, the SHELF software fits the piece-wise linear function on the transformed X^* scale to the transformed medians $m1^*$ to $m5^*$ and then transforms back to the X scale.)

Link: The final step combines the median model with the $Y3$-distribution through a link function to create the full set of conditional distributions of X given all possible values of Y. The basic approach simply defines the conditional distribution of X given $Y = y$ to be the $Y3$-distribution shifted so that its median is $m(y)$ but with otherwise the same shape. In particular, the conditional variance is constant.

(If a log or logistic transformation is used, the SHELF software again does this on the X^* scale before transforming back to the X scale.)

> *Example 2.* Remember that in this example X is the time (in days) taken to complete a building project, while Y is the average daily temperature (in degrees Celsius). In a separate elicitation workshop the "extension variable distribution" is elicited from a group of experts and is a normal distribution with mean 7 and standard deviation 6. The five recommended conditioning points

are therefore the 5th percentile $Y1 = -3$, the lower quartile $Y2 = 3$, the median $Y3 = 7$, the upper quartile $Y4 = 11$ and the 95th percentile $Y5 = 17$. (With the exception of the median, these numbers have been rounded. For instance, the 95th percentile is 16.869 but it is much easier for the experts to think about their conditional judgements if the conditioning points are rounded.)

The experts agreed on RIO judgements of the outer medians as $m1 = 88$ and $m5 = 44$.

The $Y3$-distribution is elicited as gamma with parameters 0.24 and 12, which has a median of $m3 = 48.6$. This is much closer to 44 than to 88 because the experts judged that at an average temperature of –3 there would be considerable loss of time due to freezing conditions.

The inner medians were elicited as $m2 = 64$ and $m4 = 46$.

The log transformation is indicated here. However, to illustrate the effect of the transformation, we first proceed using the identity transformation.

Figure 1 shows the basic piecewise-linear median model applied to the elicited medians. Thus, for values of Y between $Y1$ and $Y5$, the function $m(y)$ is made up of straight line segments joining the five elicited medians. For values of Y below $Y1$, the straight line between $m1$ and $m2$ is extended backwards, and similarly for values above $Y5$ the line between $m4$ and $m5$ is extended forwards.

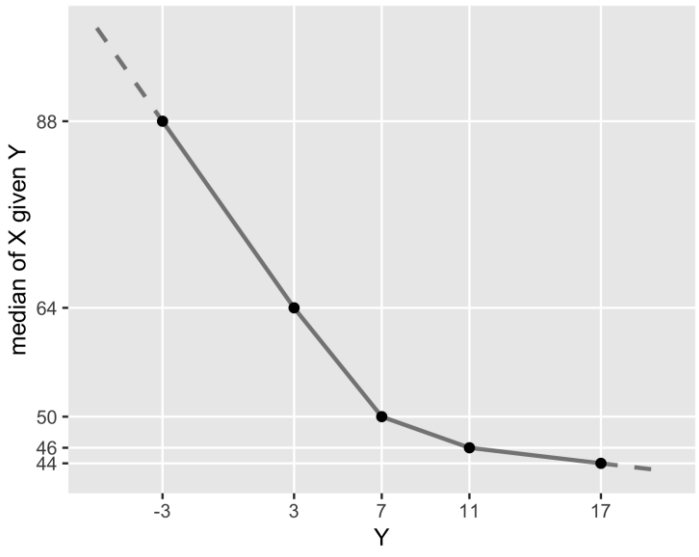

Figure 1

Finally, applying the basic link function completes the elicitation of the conditional distribution of X given Y. The distributions are plotted in Figure 2 for $Y = -3$, 7 and 17, showing how the conditional median follows the elicited values while the shape of the

distribution remains the same. In this example, the experts might feel that this link is inappropriate because they (and RIO) would have more uncertainty about the value of X conditional on $Y = -3$.

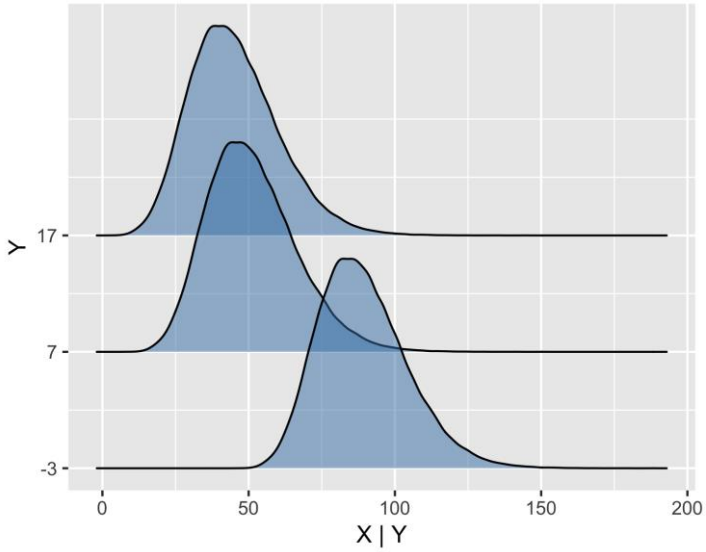

Figure 2

If instead we use the indicated log transformation, Figure 3 shows the median model. Between conditioning points, there is slight curvature, but the result is almost indistinguishable from using the identity.

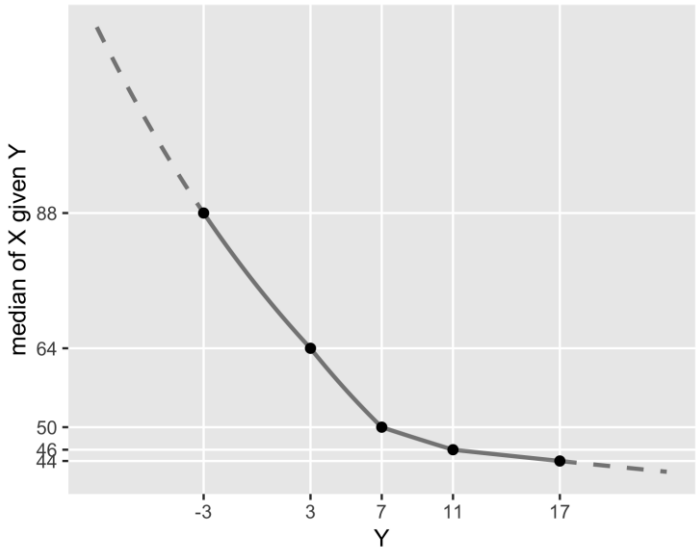

Figure 3

However, Figure 4 shows that the log transformation has a more noticeable effect on the final step. The Y3-distribution is the same as before, but now the distribution conditional on $Y1$ is appreciably more spread and that for $Y5$ is less spread. The desirability of increased

uncertainty when $Y = Y1$ was noted before, and so the experts would prefer using the log transformation here.

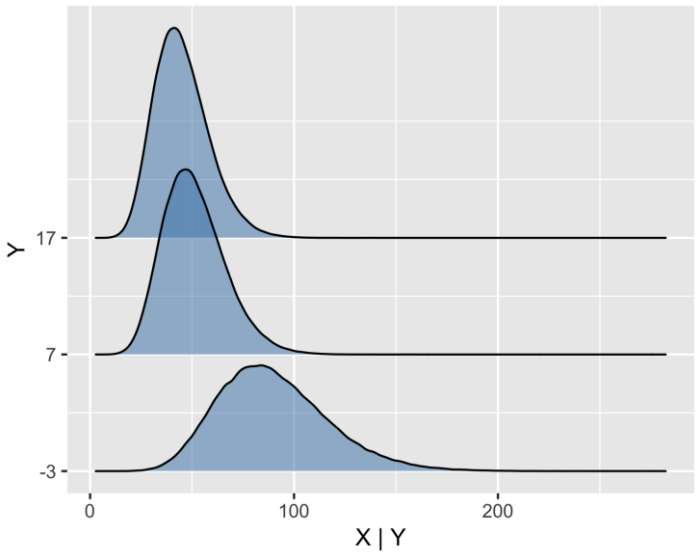

Figure 4

Variations. Although the simple basic recommendations will usually yield elicited distributions that are plausible and acceptable as RIO beliefs about X and Y, there are many alternative variations that could be used.

The recommended set of five conditioning points will generally cover the range of possible Y values well, but the facilitator may wish to use more or different points to capture how beliefs about X change with Y over particular regions of possible Y values.

Other transformations could be considered and might be appropriate in special circumstances.

The piecewise-linear median model fits exactly through the elicited points, so that $m(y)$ has the property that $m(Y1) = m1$, $m(Y2) = m2$, and so on. But these medians, like all elicited judgements, should not be treated as precise, and there may sometimes be merit in fitting a simpler median model, such as a single straight line or a quadratic. Now $m(y)$ may no longer have the property of fitting all the elicited medians exactly, but its simpler form may be preferred by the experts. Transformations of Y may also be considered for the median model.

Finally, there surely exist many other ways of building the conditional distributions of X for all possible values of Y, based on eliciting judgements conditional on a small number of values (conditioning points) of Y. Any such approach will necessarily make strong assumptions; in the recommended approach the key assumption lies in the link function, and to a lesser extent in the median model. Experience in using extension may in due course suggest new approaches.

The "Feedback" field

Feedback should be given with a view to determining whether the experts are content with the elicited distributions.

First, the conditional distributions of X for several values of Y can be plotted together to show how the distribution evolves with Y. In practice, this will usually already have been done when agreeing on a suitable link function.

Similarly, the conditional median and quartiles/tertiles might be plotted against Y. Such plots and computations help the experts to see and assess the conditional distributions. If experts express any concerns, it may be necessary to revisit the earlier judgements.

It is also important to show the implied distribution of X, particularly since this is often the primary outcome of the extension. In some special cases, it is possible for the distribution of X to have a standard distributional form. For instance, if the elicited marginal distribution of Y is normal, the $Y3$-distribution is normal, no transformation is used and the median model is a straight line, then the marginal distribution of X will also be normal. The joint distribution of X and Y is then bivariate normal.

However, in general the distribution of X cannot be written down explicitly. Instead it is implied by the other elicited distributions, and we will refer to it as *implicit*. The simplest way to compute such a distribution is by simulation.

a. Sample a large number of random values of Y from its marginal distribution. We will call them y_1, y_2, …, y_n, where n is the number of random values sampled.

b. For each sampled y_i, sample a random value of X from its conditional distribution given that $Y = y_i$. We will call this value x_i.

The resulting sample of n values of X are a simulation of the distribution of X. They can be plotted as a histogram or as a smoothed density function. The SHELF R software includes tools for doing this. Note also that the n pairs (x_i,y_i) are a simulation of the joint distribution of X and Y.

Example 2. Combining the elicited normal distribution for Y with the gamma conditional distributions of X given Y, applying the piecewise linear median model and basic link function with the log transformation, produces the marginal distribution in Figure 5. This is the outcome of the elicitation. Some summaries of this distribution are: 5th percentile 29.4, lower quartile 42.0, median 53.5, upper quartile 69.1, 95th percentile 102.8.

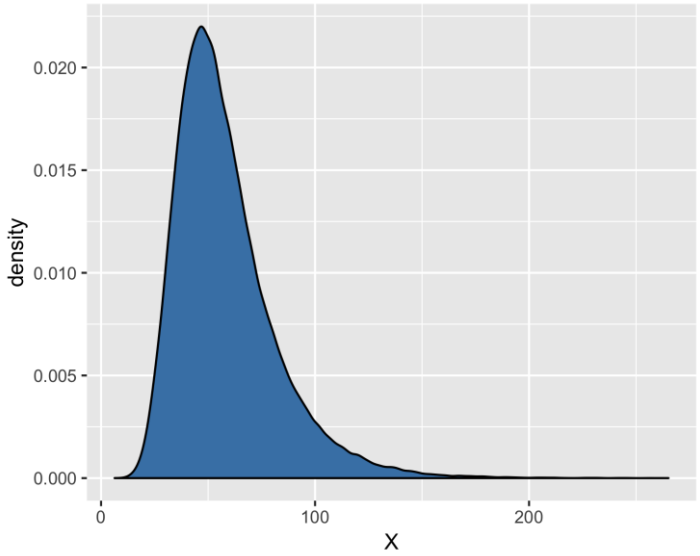

Figure 5

Case B: Discrete *X*, continuous *Y*

This is a rather uncommon situation.

Because *Y* is continuous the guidance for the "Extension variable distribution" and "Conditioning points" fields is the same as for Case A.

The "Target quantity distributions" field

When *X* is discrete, its conditional distribution given any value of *Y* comprises a probability for each of its possible values, namely the probability that *X* will take that value assuming that *Y* takes its given value.

Instead of eliciting just one conditional distribution, conditional on *Y* equalling a central conditioning point *Y*3, it is recommended to elicit the full, discrete conditional distribution of *X* for each of the conditioning points. A SHELF 3 Discrete template should be used for each of these.

Following the idea of a median model when *X* is continuous, such a model may be fitted to the elicited conditional distributions to derive probability distributions for *X* conditional on every possible value of *Y*. Since probabilities must lie between 0 and 1, the logistic transformation is indicated.

Example 2a. For the purposes of Case B we modify Example 2 by supposing that if the number of days for the project exceeds 65 the company will suffer a substantial financial penalty. We therefore redefine the target quantity to the event of the time exceeding 65, so now *X* is discrete with just two possible values. With the same conditioning points, the experts provide the following elicited conditional probabilities for the event *X*. When $Y = 17$, the probability is 0.08, it is 0.1 when $Y = 11$, 0.15 when $Y = 7$, 0.5 when $Y = 3$ and 0.85 when $Y = -1$. Figure 6 shows the result of fitting the

piece-wise linear model to these values using the logistic transformation.

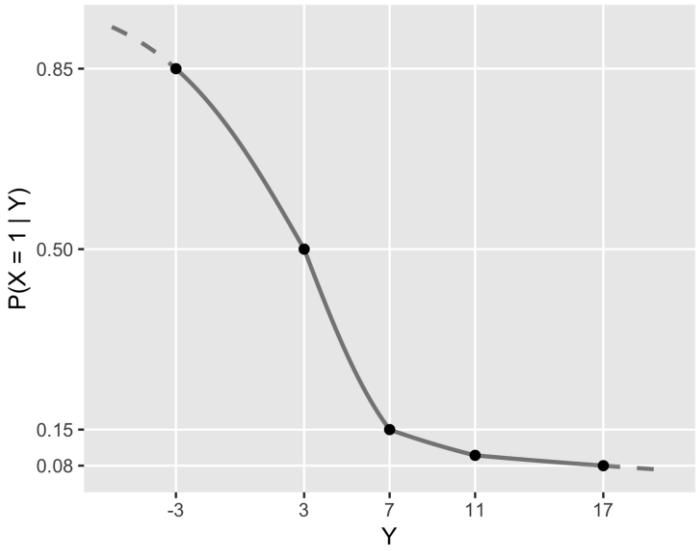

Figure 6

The "Feedback" field

Although the marginal distribution of *X* may sometimes be derived explicitly, simulation remains a simple and practical solution and would proceed as in Case A.

Example 2a. The marginal distribution in this case is simply the marginal probability of the event that the project exceeds 65 days. By making many random draws from the elicited distribution of *Y*, calculating the conditional probability of *X* from the median model at each sampled *Y* value and then averaging these, we obtain P(*X*) = 0.31.

Case C: Continuous *X*, discrete *Y*

This rather more common situation also has some similarities with Case A.

The "Extension variable distribution" field

Since *Y* is discrete, the appropriate template for this elicitation is SHELF 3 Discrete.

The "Conditioning points" field

It is recommended that the conditioning points set should comprise all the possible values of *Y*.

The "Target quantity distributions" field

If the number of possible values of *Y* is sufficiently small, it will be feasible to elicit the joint distribution of *X* conditional on each of the conditioning

points (and hence on every possible Y value), completing a separate SHELF 2 template for each.

Alternatively, we may proceed as in Case A. Thus, a conditional distribution is elicited for a single central conditioning point and medians elicited for the other conditioning points. It is not necessary to create a median model, since we will have the median for every possible value of Y, but the basic link function should be used (with transformation as appropriate) to derive the full set of conditional distributions.

The "Feedback" field

The marginal distribution of X is now a simple mixture of the conditional distributions and may be plotted explicitly.

Example 3. Here, X is the rainfall in the next year in South Africa, in millimetres averaged over the country. Y is the discrete extension variable taking values El Niño, Neutral and La Niña, so these form the three conditioning points. The extension variable distribution is elicited as having probabilities P(El Niño) = 0.05, P(Neutral) = 0.55 and P(La Niña) = 0.4.

The target quantity distributions at the three conditioning points were elicited as follows:

In an El Niño year, gamma with parameters 0.03, 9.3 (median 300).

In a Neutral year, gamma with parameters 0.011, 5.8 (median 500).

In a La Niña year, gamma with parameters 0.03, 24.3 (median 800).

These are shown in the Figure 7 below, together with the resulting distribution of X (a weighted average of the three gamma distributions with weights given by the distribution of Y).

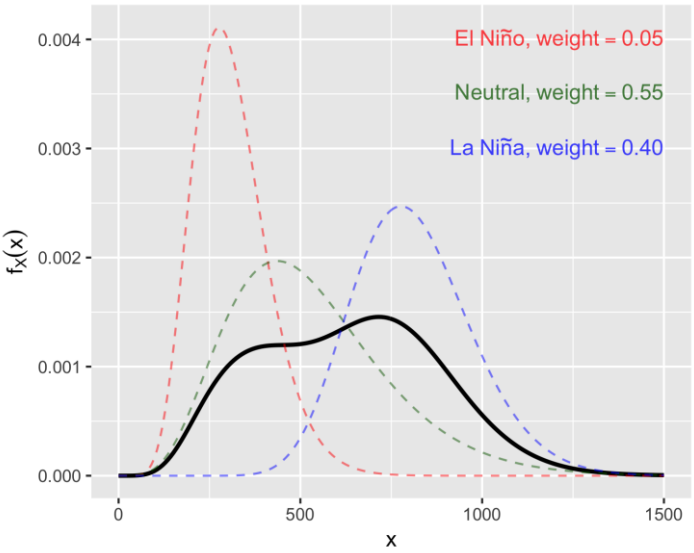

Figure 7

The marginal distribution of rainfall in Figure 7 shows a hump around 300-500mm and another around 800mm. A distribution with two separate humps is called bimodal, and such distributions may easily arise in Case C. (In this example the humps are not quite separated, so the distribution is technically still unimodal.)

Case D: Discrete *X*, discrete *Y*

The guidance for this case is particularly simple.

The "Extension variable distribution" field

Since Y is discrete, the appropriate template for this elicitation is SHELF 3 Discrete. The resulting distribution gives the probability $P(Y = y_i)$ for each possible value y_i of Y.

The "Conditioning points" field

It is recommended that the conditioning points set should comprise all the possible values of Y.

The "Target quantity distributions" field

For each conditioning point, i.e. for each y_i, the discrete conditional distribution of X given that Y equals y_i is elicited using the SHELF 3 Discrete template. Thus, for each possible value x_j of X and for each y_i, we have the conditional probability $P(X = x_j \mid Y = y_i)$.

The "Feedback" field

The joint distribution of X and Y now comprises the probabilities

$$P(X = x_j , Y = y_i) = P(X = x_j \mid Y = y_i) \, . \, P(Y = y_i)$$

and the marginal probability that X equals x_j is obtained by summing these joint probabilities over all the possible y_i values, i.e.

$$P(X = x_j) = \Sigma_i \, P(X = x_j , Y = y_i) \, .$$

Example 4. In this example, X is the result of a urine test for kidney function, while Y is the event that the patient has kidney disease. The test result is a measurement known as GFR, which is a continuous quantity, and if we consider the measured GFR to be the target quantity then we are in Case C. However, in this example we let X be discrete. The GFR for the majority of healthy adults is above 80, while a value below 60 suggests chronic kidney disease. We will define X to be the GFR result reported according to these boundaries as Normal (N), Moderate (M) or Low (L).

From symptoms found when examining the patient, the doctor judges the probability of kidney disease in this patient to be P(Disease) = 0.3, and therefore P(Healthy) = 0.7. This is the elicited extension variable distribution.

The doctor then judges the target variable's distribution for a healthy patient as comprising

P(N | Healthy) = 0.62, P(M | Healthy) = 0.30, P(L | Healthy) = 0.08.

Finally, the patient's symptoms do not suggest very serious kidney disease, and the doctor judges the following probabilities for the target variable if the patient does have kidney disease:

P(N | Disease) = 0.05, P(M | Disease) = 0.10, P(L | Disease) = 0.85.

The elicited joint distribution for X and Y is now given by the following table.

X / *Y*	Normal	Moderate	Low	*Total*
Disease	0.015	0.030	0.255	*0.3*
Healthy	0.434	0.210	0.056	*0.7*
Total	*0.449*	*0.240*	*0.311*	*1.0*

The bottom row is the marginal distribution of X. However, in this example the primary objective is not this distribution or the joint distribution shown in the whole table. Instead, the doctor will need to assess the probability of the patient having kidney disease after receiving the test result. For instance, if the result is Normal, then

P(Disease | N) = P(Disease and N) / P(N) = 0.015/0.449 = 0.033.

If the test result is Moderate or Low the doctor's probability of the patient having kidney disease is

P(Disease | M) = 0.03/0.24 = 0.125,

P(Disease | L) = 0.255/0.311 = 0.820.

Therefore, if the GFR result is Low (below 0.6) the doctor will have an 82% probability for the patient having kidney disease, and can prescribe suitable treatment. If the result is Normal or Moderate, the doctor should consider other sources of the patient's symptoms.

Multivariate quantities

Either or both of X and Y could be multivariate, composed of two or more uncertain quantities rather than just one. The extension approach can be applied to such cases in principle but in practice will be more complex.

If Y is multivariate, then the "Extension variable" field will still require the distribution of Y to be elicited and recorded, but now that requires multivariate elicitation which represents a more complex elicitation challenge. Furthermore, the median model $m(y)$ is now a function of two or more variables which poses additional complications. There will certainly be a need for more conditioning points.

If X is multivariate, its conditional distribution given any value of Y is multivariate. There is no longer a single median and $m(y)$ becomes a vector-valued function.

Techniques to address these challenges will depend on context, but one useful case can be identified and dealt with relatively simply.

Chained quantities. Consider Example 1, where interest lies in the sales of a new product in its first two years. The company might of course be interested in sales over the first three years. If we denote sales in years 1, 2 and 3 as $S1$, $S2$ and $S3$, then we can begin by using Extension with $S3$ as the target variable and ($S2$, $S1$) as the (multivariate) extension variable. We will refer to this as the $S3$-extension. Experts may judge that the conditional distribution of $S3$ given both $S2$ and $S1$ will not depend on $S1$. That is, as long as the experts know the value of second year sales, then knowing the first year sales does not help to predict sales in the third year. Then throughout the "Target quantity distributions" stage we can treat $S2$ as the extension variable because we only need values of $S2$ as conditioning points.

We will still need to elicit the multivariate distribution of $S2$ and $S1$ as the extension variable distribution, but this can be done with another SHELF 3 Extension template, in which $S2$ is now the target quantity and $S1$ is the extension variable. We will refer to this as the $S2$-extension. The distribution of $S2$ and $S1$ will now typically be implicit, represented by a sample of pairs of values (x_i, y_i). We will need to record the 5^{th} and 95^{th} percentiles, median and quartiles of the x_i values in the $S2$-extension because these will be needed as conditioning points in the $S3$-extension. The full set of x_i values are also taken forward as sampled values of $S2$ for computing feedback in the $S3$-extension (where they are now regarded as y_i values). The completed SHELF 3 Extension template in the Samples folder of the SHELF package shows the $S2$-extension for Example 1.

Whenever we have a collection of three or more quantities of interest that can be put into a sequence such that each member of the sequence depends on the immediately preceding quantity but not on earlier ones, then we can elicit their distributions using a chain of extensions.