Appendix. Mathematical details of model

A.1 Discrete probability density function of the number of food servings per week

The number of servings per week of each food group is characterised by a probability density function (PDF). Because the number of servings is an integer variable, a discrete PDF is more appropriate than a continuous PDF. The formulation of the PDF is informed by empirical data on the minimum, median and maximum number of servings per week of each food group.

Figure 2 shows the discrete PDF which is assumed to have a triangular-like (a two-split straight lines) pattern. The PDF is characterised by six parameters (a, c, b, l, u, m) which are food-group specific. The parameters a, c and b are assumed to be known; they are based on the observed food consumption patterns of the population of interest. They are respectively the lower bound, median and upper bound for the number of food group servings per week that can occur in diets consumed by the target population. The parameters l, u and m on the other hand are the unknown probabilities to be determined. They correspond respectively to the probabilities of the lower bound, median and upper bound for the number of servings from a specific food group in any diet.

The probabilities (l, u, m) should satisfy a number of constraints. In general there is no unique solution (l, u, m) for any given (a, c, b) , and in some cases there is no solution. The calculation of the probabilities is performed in two steps: the first step formulates the set of constraints (equalities and inequalities) that should be satisfied by the probabilities, and the second step calculates the probabilities by solving simultaneously the set of equalities and inequalities.

<u>Step 1</u>

The probabilities must satisfy a number of equality and inequality constraints. The inequality constraints are (Figure 2):

$$
\begin{gathered}
l > 0 \\
u > 0 \\
m > l \\
m > u
\end{gathered}
\tag{1}
$$

There are two equality constraints:

$$\sum_{k=a}^{c} p_k = \frac{1}{2} \tag{2}$$

$$\sum_{k=a}^{b} p_k = 1 \tag{3}$$

where p_k is the probability that the diet contains k servings per week of the food group. The set of inequalities in (1) ensure that the probabilities are positive and the probability of the median should be higher than the probability of either extreme. Equation (2) is the definition of the median and Equation (3) ensures that the probabilities over the range of servings add up to one. Because the PDF is assumed to have the two-split straight line pattern shown in Figure (2), p_k is defined as:

$$\mathrm{p_k} = \begin{cases} l + (m - l)\dfrac{k - a}{c - a}, & \text{if } a \le k \le c \\ u + (m - u)\dfrac{k - b}{c - b}, & \text{if } c < k \le b \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Combining Equation (2) and the first equation in (4) gives

$$\sum_{\mathrm{k=a}}^{\mathrm{c}} \mathrm{p_k} = \frac{1}{2} = l + \sum_{k=a+1}^{c-1} \left(l + (\mathrm{m} - l)\left(\frac{k - a}{c - a}\right)\right) + m$$

$$= l + \left(\sum_{k=a+1}^{c-1} l \right) + \left(\left(\frac{m-l}{c-a} \right) \sum_{k=a+1}^{c-1} (k-a) \right) + m$$

$$\frac{1}{2} = l + l(c-a-1) + \left(\left(\frac{m-l}{c-a} \right) \sum_{k=1}^{c-a-1} k \right) + m$$

(5)

Noting that

$$\sum_{k=1}^{c-a-1} k = \frac{(c-a-1)(c-a)}{2}$$

(6)

and substituting Equation (6) on the right hand side of Equation (5) gives:

$$\frac{1}{2} = l(c-a) + \frac{(m-l)(c-a-1)(c-a)}{2(c-a)} + m$$

$$\frac{1}{2} = l(c-a) - \frac{l}{2}(c-a-1) + \frac{m}{2}(c-a-1) + m$$

$$\frac{1}{2} = \frac{l}{2}(c-a+1) + \frac{m}{2}(c-a+1) = \frac{(l+m)}{2}(c-a+1)$$

(7)

But

$$\sum_{k=a}^{b} p_k = 1 = \sum_{k=a}^{c} p_k + \sum_{k=c+1}^{b-1} p_k + u$$

(8)

Combining Equations (3), the second equation in (4), (7) and (8) gives:

$$\sum_{k=a}^{b} p_k = 1 = \sum_{k=a}^{c} p_k + \sum_{k=c+1}^{b-1}\left(u + (m-u)\left(\frac{k-b}{c-b}\right)\right) + u$$

$$1 = (l+m)\frac{(c-a+1)}{2} + u(b-c-1) + \left(\left(\frac{m-u}{c-b}\right)\sum_{k=c+1}^{b-1}(k-b)\right) + u$$

$$1 = (l+m)\frac{(c-a+1)}{2} + u(b-c) + \frac{m-u}{(c-b)}h$$

$$1 = \frac{l}{2}(c-a+1) + m\left(\frac{1}{2}(c-a+1) + \frac{h}{c-b}\right) + u\left(b-c-\frac{h}{c-b}\right)$$

(9)

where

$$h = \sum_{k=c+1}^{b-1}(k-b) = \frac{(b-1)b}{2} - \frac{c(c+1)}{2} - b(b-c-1)$$

(10)

Finally, combining Equations (7) and (9) gives

$$\begin{pmatrix} e & e & 0 \\ e & f & g \end{pmatrix}\begin{pmatrix} l \\ m \\ u \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix}$$

(11)

where

$$e = \frac{c-a+1}{2}$$

$$f = e + \frac{h}{c-b}$$

$$g = b - c - \frac{h}{c-b}$$

(12)

In other words, (l, u, m) should satisfy the inequalities in (1) and Equation (11).

Step 2

There is no unique solution to three unknowns (l, u, m) which satisfy (1) and (11). One solution can be obtained by solving the following linear programming (LP) problem:

LP (1)

$$\max_{(l,u,m)} \{m\}$$
$$such\ that$$
$$\begin{pmatrix} e & e & 0 \\ e & f & g \end{pmatrix} \begin{pmatrix} l \\ m \\ u \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix}$$
$$l > 0$$
$$u > 0$$
$$m > l$$
$$m > u$$

(13)

The solution of LP (1) will ensure that (1) and (11) are satisfied whilst maximising m. Equally a solution of (1) and (11) can also be obtained by solving a similar LP problem but where the objective is to minimise m instead of maximising m:

LP (2)

$$\min_{(l,u,m)} \{m\}$$
$$such\ that$$
$$\begin{pmatrix} e & e & 0 \\ e & f & g \end{pmatrix} \begin{pmatrix} l \\ m \\ u \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix}$$
$$l > 0$$
$$u > 0$$
$$m > l$$
$$m > u$$

(14)

It makes sense to average the solutions so that they are not biased to either extreme of m:

$$\widetilde{m} = \left(\frac{m_1^* + m_2^*}{2}\right)$$
$$\tilde{l} = \left(\frac{l_1^* + l_2^*}{2}\right)$$
$$\tilde{u} = \left(\frac{u_1^* + u_2^*}{2}\right)$$

(15)

where (m_1^*, l_1^*, u_1^*) and (m_2^*, l_2^*, u_2^*) are respectively the solution sets of LP (1) and LP (2) problems. Because the constraints are linear and convex, the solution obtained by averaging the solutions of LP (1) and LP (2) will also satisfy the constraints.

It is also possible to choose nontrivial a, c and b such that Equations (2) and (3) cannot be satisfied. One example is the triplet: $a = 1$, $c = 2$ and $b = 10$. For this example, no solution can be obtained for the probabilities (l, m, u). In order to solve for (l, m, u) in this example (and other similar examples) the following algorithm is used to move the values of c and b in a consistent way:

1. For the trivial case $a = b = c$, set $l = m = u = 1$.
2. Define M to be the maximum allowable movement, defined here as $M = b/4$ (rounded up to the nearest integer).
3. Set $\acute{b} = b, \acute{c} = c$.
4. Let $b = b + 1$.

4.1. If $b - \acute{b} < M$, attempt to solve the LP problem again. If successful stop, otherwise return to Step 4.

4.2. If $b - \acute{b} \geq M$, let $b = \acute{b}$, and proceed to Step 5.

5. Let $c = c + 1$

5.1. If $c - \acute{c} < M$, attempt to solve the LP problems again. If successful stop, otherwise return to Step 4.

5.2. If $c - \acute{c} \geq M$, then no solution can be found.

This algorithm attempts to move b, and then c, in a positive direction to obtain soluble LP problems. If M is small, it may not succeed.

A.2 Energy and nutrient contents of a random diet

Recall that the number of servings per week of a food group is described by the PDF shown in Figure 2 and that this number is distributed amongst the food items within the food group according to another discrete PDF.

For each nutrient k the total nutrient content in diet i is given by:

$$N_{ik} = \sum_j x_{ij} s_j h_{jk} \tag{16}$$

where x_{ij} is the number of servings of food item j consumed in random diet i, s_j (a constant) is the average weight in grams of a serving of food item j, and h_{jk} is the amount of nutrient k per gram in food item j. Similarly, the energy content of diet i is:

$$E_i = \sum_j x_{ij} s_j q_j \tag{17}$$

where q_j be the amount of energy per gram in food item j.

A.3 Scaling portion sizes

To ensure that the distribution of the energy content of the baseline diet corresponds with the distribution of the energy requirement of the population of interest, the portion sizes of the food items are scaled. In practice it is sufficient to ensure that the first two moments of the simulated energy content are as close as possible to those of the population energy requirements. The energy content of the random diet i can be re-written as

$$E_i = \delta_i \sum_j x_{ij} s_j q_j \tag{18}$$

Equation (18) is the same as Equation (17) if $\delta_i = 1$. The random scalar variable δ_i is introduced to vary the portion size. For simplicity it is assumed to be a common multiplicative dimensionless scalar by which to adjust all portion sizes of all food items such that mean $\bar{E}$ and variance $\tilde{E}$ of E_i are identical (or as close as possible) to the mean and variance of the required energy diet content, θ and φ. We will use henceforth the bar symbol over a variable to indicate the mean of the variable and the tilde symbol to indicate its variance.

Without loss of generality we can assume each δ_i to be a uniformly distributed positive random variable between $\bar{\delta} - 3\tilde{\delta}$ and $\bar{\delta} + 3\tilde{\delta}$, where $\bar{\delta}$ and $\tilde{\delta}$ are given. Because δ_i and x_{ij} are independent, then from Equation (18):

$$\bar{E} = \bar{\delta} \sum_j \bar{x}_j s_j q_j \tag{19}$$

where $\bar{x}_j$ is the mean number of servings per week of food item j. For $\bar{E} = \theta$ (required mean of energy intake), it is necessary and sufficient that

$$\bar{\delta} = \frac{\theta}{\sum_j \bar{x}_j s_j q_j}$$

(20)

Now denote by

$$y_i = \sum_j x_{ij}\, s_j q_j$$

(21)

Substituting y_i in Equation (18) gives

$$E_i = \delta_i\, y_i$$

(22)

Because δ_i and y_i are independent then the variance of the energy content of the diet is given by

$$\tilde{E} = \left(\tilde{\delta} + \bar{\delta}^2\right)(\tilde{y} + \bar{y}^2) - \bar{\delta}^2\, \bar{y}^2$$

(23)

From Equation (23), for $\tilde{E} = \varphi$ (the required variance of energy content) it is necessary and sufficient that

$$\tilde{\delta} = \frac{\varphi - \bar{\delta}^2 \tilde{y}}{\bar{y}^2 + \tilde{y}}$$

(24)

Note that Equation (24) will have a feasible solution if and only if

$$\varphi \geq \bar{\delta}^2 \tilde{y}$$

(25)

If inequality (25) does not hold this means that the variability in the model is less than that in the data observations and the PDFs should be revised.

Assuming that inequality (25) holds, $\tilde{\delta}$ can be calculated by substituting Equations (26) and (27) below − which are determined directly from Equation (21) − into Equation (24):

$$\tilde{y} = \sum_j \tilde{x}_j \, s_j^2 \, q_j^2 \tag{26}$$

and

$$\bar{y}^2 = \left(\sum_j \bar{x}_j \, s_j \, q_j \right)^2 \tag{27}$$

The variable nutrient intake of the set of recommendations is then calculated by:

$$N_{ik} = \delta_i \sum_j x_{ij} \, s_j h_{jk} \tag{28}$$