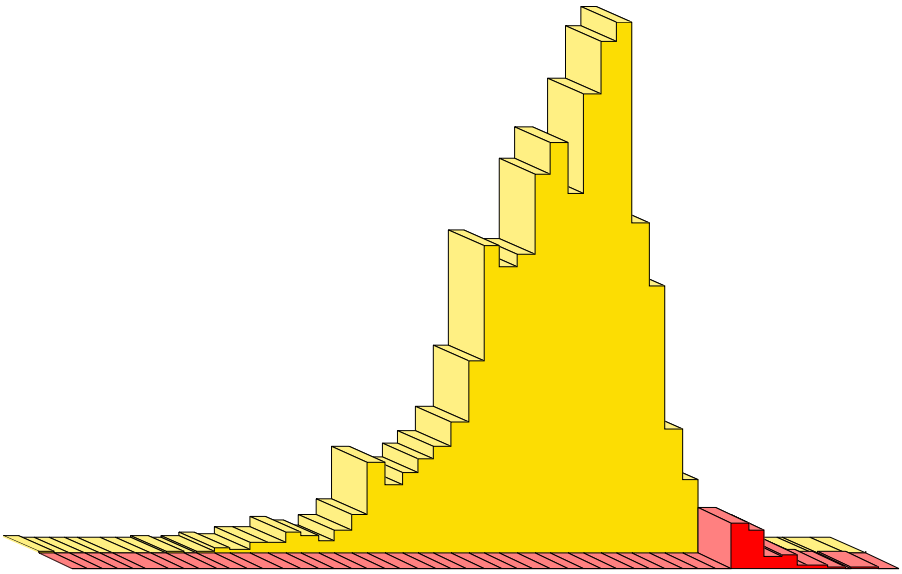
MCRA

a web-based program for Monte Carlo Risk Assessment

Release 3

Reference Guide

Hilko van der Voet
Waldo J. de Boer
Polly E. Boon
Gerda van Donkersgoed
Jacob D. van Klaveren



Biometris and RIKILT
Wageningen University and Research centre

March 2004

Biometris is the integration of the Centre for Biometry of Plant Research International and the Department of Mathematical and Statistical Methods of Wageningen University and is a part of Wageningen University & Research centre.

Post address
P.O. Box 100
6700 AC Wageningen
The Netherlands
Visiting address
Bornsesteeg 47, building no. 116
6708 PD Wageningen
Telephone: +31 (0)317 476925
Telefax: +31 (0)317 483554

RIKILT (Institute of Food Safety) is a part of Wageningen University & Research centre.

Post address
P.O. Box 230
6700 AE Wageningen
The Netherlands
Visiting address
Bornsesteeg 45, building no. 123
6708 PD Wageningen
Telephone: 0031 317 475400
Telefax: 0031 317 417717

Contents

1 Foreword

This Reference Manual describes a stochastic (or Monte Carlo) model for dietary exposure assessment of chemical substances based on monitoring data concerning the quality of agricultural products. Exposure assessment is an important step in the risk assessment of chemical substances, such as agricultural chemicals (pesticides, veterinary drugs), toxins (e.g. mycotoxins) and environmental contaminants (e.g. dioxins).

The methods for probabilistic modelling described here are implemented in the program MCRA (Monte Carlo Risk Assessment). MCRA can calculate intake distributions for both short-term (acute) and long-term (chronic) exposures. Basically, for an assessment of acute risks MCRA simulates daily consumptions by sampling a food consumption database, and combines these with a random sample from either a residue database (empirical distribution) or a parametric distribution of residue levels. The result is a full *distribution* of intakes, rather than traditional deterministic methods which only provide a point estimate. Percentiles of the intake distribution can be used to assess risks by relating them to e.g. an acute reference dose (ARfD). In a chronic risk assessment, MCRA calculates the distribution of the usual exposure based on the average residue level and the empirical distribution of consumption between individuals and between different consumption days of the same individuals. Percentiles of this usual intake distribution can then be related to e.g. the acceptable daily intake (ADI).

Uncertainty of percentiles can be established by bootstrapping. It is possible to include processing factors (e.g. the effect of cooking on the residue level) and variability factors (to correct for the fact that monitoring data are obtained from composite samples, whereas consumers may eat individual units). Analyses can be done for a total population or for a subpopulation (e.g. children, or consumption-days only). The effects of residue levels below analytical reporting limits can be assessed.

The program MCRA is a result of an ongoing co-operation between RIKILT and Biometris since 1998. RIKILT co-ordinates the Dutch KAP programme (Quality of Agricultural Products) where results of monitoring programs for chemical residues in food are gathered in a national database. RIKILT also has a recipe database to link food codes from the Dutch food consumption table to primary agricultural products. Biometris contributes statistical models and programs for quantitative risk analysis.

This Reference Guide gives a complete description and justification of the statistical methods used in the program MCRA. For a practical introduction into a dietary exposure assessment with MCRA please consult the MCRA User Manual (de Boer *et al*., 2004).

2 MCRA, an introduction

This documentation gives a description and justification of the statistical methods implemented in the program MCRA (Monte Carlo Risk Assessment).

MCRA can be used for assessment of acute and/or chronic risks due to the intake of residues on food. MCRA provides the following options:

- acute risk assessment
- chronic risk assessment
- parametric or non-parametric modelling of residue levels
- modelling of processing effects
- modelling of sample variability
- modelling of nondetects levels
- restrictions on age and/or consumption days
- calculate exposure distribution for consumption-days only
- bootstrap sampling of consumers and/or processing factors to assess the uncertainty of percentiles
- deterministic estimates (IESTI)

For a practical application of a dietary exposure assessment with the program MCRA we refer to the MCRA User Manual (de Boer *et al*., 2004). In the next chapters, the theory of probabilistic modelling with MCRA is described.

3 Model description

3.1 Basic model

This chapter describes the stochastic (or Monte Carlo) models behind the MCRA program. These models assess acute or chronic risks due to the intake of chemical substances from food by combining food consumption survey data and residue concentration data from e.g. monitoring programs. The model for acute risk allows for effects of food processing between monitoring and ingestion, it can model unit variability either from available data or using default assumptions, and it uses information on Limit of Reporting (LOR) and percent crop treated to check whether nondetects present a source of uncertainty.
The basic model is:

$$y_{ij} = \frac{\sum_{k=1}^{p} x_{ijk}\ c_{ijk}}{w_i}$$

where y_{ij} is the intake by individual i on day j (in µg chemical substance per kg body weight), x_{ijk} is the consumption by individual i on day j of food commodity k (in g), c_{ijk} is the concentration of the chemical substance in commodity k eaten by individual i on day j (in mg/kg, 'ppm'), and w_i is the body weight of individual i (in kg). Finally, p is the number of food commodities accounted for in the model. Note that the definition of 'commodity' is flexible: it may represent a Raw Agricultural Commodity (RAC), e.g. 'apple', but it may also specify subdivisions, e.g. 'apple, peeled' or 'apple, imported'.

In the stochastic model the quantities x_{ijk}, w_i and c_{ijk} are assumed to arise from probability distributions describing the variability for individual food consumption and weight, $p(x_1,...,x_p,w)$, and for residue concentrations in each food commodity, $p_k(c)$. In principle, these probability distributions may be parametric (e.g. completely defined by the specification of some parameter values) or empirical (e.g. only implicitly defined by the availability of a representative sample).

Food consumption data may arise from different sources. Typically national food consumption surveys or monitoring programs provide information on food intake in the general population. For example, from the Dutch Food Consumption Survey 1997 food consumption patterns ($x_1,...,x_p$), body weight (w) and age (a) are available for 6250 individual persons on 2 consecutive days. Depending on the problem, Monte Carlo samples may be drawn from the complete data base, from a day- or age-restricted subset or from consumption-days only. In some cases there is insufficient information for specific subgroups in the population. For example, in a study on infants (age up to 12 months), a separately constructed food consumption database has been used (Boon *et al.* 2003).

In general a recipe data base is necessary to convert the amounts of food as consumed (*e.g.* pizza) to amounts of commodities ($x_1,...,x_p$) of raw agricultural products which are used in the model. Van Dooren et al. (1995) provide such a conversion for the Dutch situation.

Residue concentration data may be available from different sources. In some countries national monitoring databases exist, which are useful for the risk assessment of chemical compounds already in use. For example the Dutch KAP database (van Klaveren 1999) stores annually more than 200.000 records of measurements originating from food monitoring programs for meat, fish, dairy products, vegetables and fruit.

Given these probability distributions (or estimates thereof) Monte Carlo simulations can be used to generate an estimate of the probability distribution $p(y_{ij})$ to assess acute risks by intake of the residue

(see 3.2). When dietary components are consumed on a nearly daily basis, intake values y_{ij} may be used to estimate the probability distribution $p(y_{i.})$ for chronic risk assessment purposes (see 3.6).

3.2 Modelling of residu concentrations in consumed food

3.2.1 Distributional assumptions

Residue concentrations in the various food commodities are independent and therefore can be modelled by univariate distributions.

3.2.1.1 Non-parametric modelling of residue levels

In the non-parametric (empirical) approach, residue values are sampled at random from the available data and combined with the consumption data to generate a new distribution of exposure values. To assess the risk-exposure, percentiles of the exposure distribution are estimated.

3.2.1.2 Parametric modelling of residue levels

In the parametric approach, residue concentrations per food commodity are sampled from parametric distributions. A special feature of residue data is that the large majority of measured concentrations (often more than 80%) is recorded as zero (nondetects). These values may correspond to true zero concentrations (for example because the substance is never used in the specific product), or they may correspond to low concentrations which are below a pre-established reporting limit (LOR). In any case, the residue concentration distribution is very skew, with a large spike at zero and an extended tail to higher values. For statistical modelling a two-step procedure is chosen. First, the presence of a concentration ≥ LOR on food products is modelled with a binomial distribution with a parameter p representing the probability of a reported residue level. Probability p depends on the chemical substance and the product and is estimated as the fraction of detects. Secondly, the non-zero residues are modelled with a parametric distribution. After consideration of several possibilities using the program *BestFit*, the lognormal distribution has been selected as being both theoretically sensible and practically useful. The parameters μ and σ are the mean and standard deviation of the log-transformed non-zero residue concentrations.
In the basic model (see 3.1)

$$c_{ijk} = I_{ijk} \cdot cpos_{ijk}$$

with I_{ijk} indicating whether a residue concentration is sampled (I_{ijk}=1) or not (I_{ijk}=0), and $cpos_{ijk}$ the residue concentration in the subpopulation of positive values. The probability of I_{ijk} being 1 or 0 depends on the number of detects found for commodity k and I_{ijk} is sampled separately for each individual i on occasion j.

3.2.2 Modelling of processing effects

Concentrations in the consumed food may be different from concentrations in the product as measured in monitoring programs (typically raw product) due to processing, such as peeling, washing, cooking etc.
In general, we assume the model:

$$cpos_{ijk} = f_k \cdot cr_{ijk}$$

where cr_{ijk} is the concentration in the raw product, and where f_k is a factor for a specific combination k of RAC and processing. Values will typically be between 0 and 1, although occasionally the processing factor may also be >1 (e.g. drying as applied for grapes and figs).
The user of the model will have to specify processing factors for each commodity k as defined in the food consumption data base. For this purpose, it is advised to maintain a data base of processing

factors, indexed by chemical substance, RAC and processing type (e.g. washing, peeling or other processing). Before running the model, it may then be necessary to specify how the necessary processing factors are derived from the data base entries and/or other information. Example: if there are no processing factors known for captan in pears, it may be decided to use the corresponding factors for apples instead.

Often the information will be of limited quality, and this may be entered in the Monte Carlo modelling by specification of uncertainties. A practical proposal is to specify for each processing factor two values:

1. $f_{k,nom}$: the nominal value, typically some sort of mean from an experimental study
2. $f_{k,upp}$: an upper 95% confidence limit, which typically will be set by an expert (even if statistical information on variability of the factor is available, there will often be uncertainty due to the appropriateness of the processing study for the population of the risk assessment). The upper limit should be such that experts will easily agree that it is not set too low.

A typical data base entry might thus read:

RAC	processing	$f_{k,nom}$	$f_{k,upp}$
apple	washing	0.5	0.7

and, confronted with the need to have processing factors for pears in a specific risk assessment, an expert may decide upon:

RAC	processing	$f_{k,nom}$	$f_{k,upp}$
pear	washing	0.5	0.8

In the Monte Carlo modelling, processing factors can be used in either of three ways (for each commodity k to be chosen by the user):

1. (no processing factor) Just take $f_k = 1$. This is in most (though not all) cases a worst-case assumption. No data on processing are needed and therefore this route is useful in a first tier approach.
2. (fixed value) Use $f_k = f_{k,upp}$. Available information on specific processing effects is used, although still in a cautionary way (in accordance with the precautionary principle). Note that $f_{k,nom}$ values need not to be specified.
3. (distribution based) Sample f_k using a normal distribution. Log or logit transformed values of $f_{k,nom}$ and $f_{k,upp}$ are used to define the first two moments of the normal distribution. Two situations are distinguished depending on the type of transformation.
   a) The logarithms of $f_{k,nom}$ and $f_{k,upp}$ are equated to the mean and the 95% one-sided upper confidence limit of a normal distribution. This normal distribution thus is specified by a mean $\ln(f_{k,nom})$ and a standard deviation $\{\ln(f_{k,upp}) - \ln(f_{k,nom})\}/1.645$. Values are drawn from this distribution in the Monte Carlo simulations. Processing factors f_k will be nonnegative. Note: $f_{k,upp}$ and $f_{k,nom}$ values equal to 0 are replaced by a low user-specified value (e.g. 0.01); this is useful computationally to avoid problems with logarithms.
   b) The logits of $f_{k,nom}$ and $f_{k,upp}$ are equated to the mean and the 95% one-sided upper confidence limit of a normal distribution. This normal distribution thus is specified by a mean $\text{logit}(f_{k,nom})$ and a standard deviation $\{\text{logit}(f_{k,upp}) - \text{logit}(f_{k,nom})\}/1.645$. Values are drawn from this distribution in the Monte Carlo simulations. Processing factors f_k will be between 0 and 1. Note: $f_{k,upp}$ and $f_{k,nom}$ values equal to 0 and 1 are replaced by user-specified values (e.g. 0.01 and 0.99); this is useful computationally to avoid problems with logits.

   The user should keep in mind that, in case of a lognormal distribution, $f_{k,nom}$ defines the median, while $f_{k,upp}$ quantifies skewness. The same holds for the logistic. Usually, a logarithm will be the standard transformation, but for very skew distributions (see Figure 1) occasionally values above 1 are sampled (upper row, 1rst, 3rd and 5th plot). A logit transformation should be considered instead.

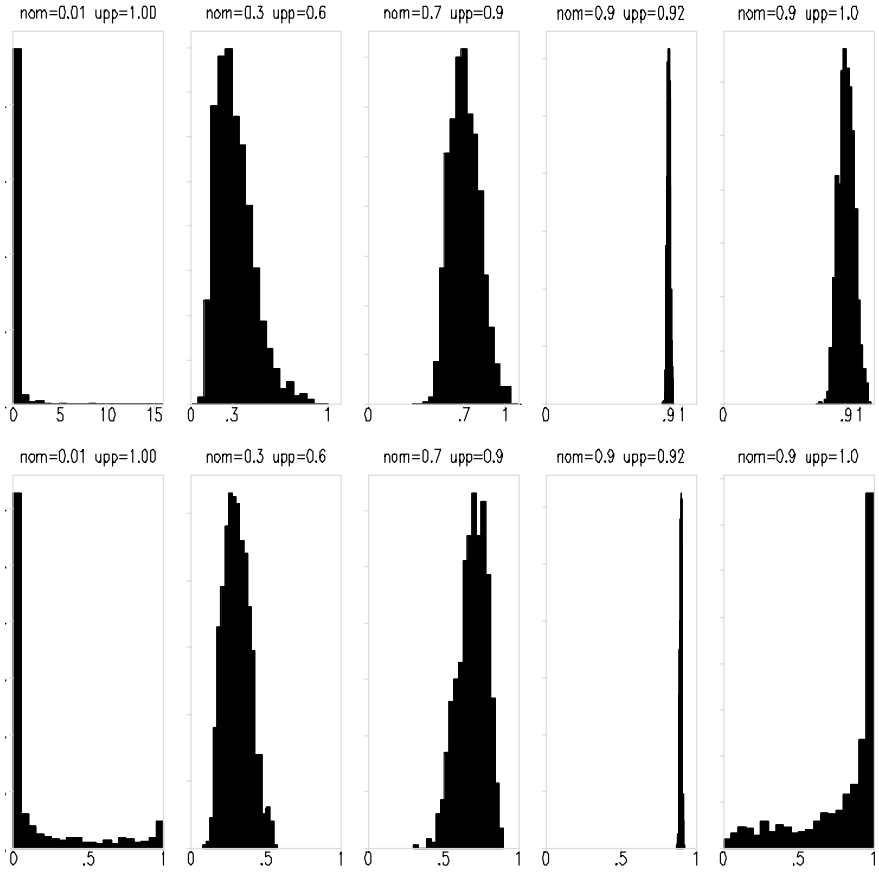


Figure 1: Lognormal (upper row) and logistic (lower row) distributions for various values of $f_{k,nom}$ (= nom) and $f_{k,upp}$ (= upp)

3.2.3 Modelling of unit variability

3.2.3.1 Introduction, variability in deterministic modelling

Variability in residue concentrations between individual units is a relevant factor in the assessment of short-term dietary exposure to chemical residues. It is addressed separately because monitoring measurements cm_k are typically made on homogenised composite samples, both in controlled field trials and in food monitoring programs. Such a composite sample for product k is composed of nu_k units with nominal unit weight wu_k each. The weight of a composite sample is therefore $wm_k = nu_k \times wu_k$. This weight is often larger than a consumer portion, e.g. a typical composite sample of 20 sweet peppers weighs 3.2 kg, whereas daily consumer portion weights in the Dutch Food Consumption Survey 1997 ranged from 0.08 g to 458 g.

How should monitoring data be used to estimate the raw commodity concentration levels cr_{ijk} in consumer portions? Although the mean level of cm_k may be a fair estimate of the mean level of cr_{ijk}, the variability of cm_k is not appropriate to estimate the variability of cr_{ijk}. In smaller portions more extreme values may occur more readily, and thus acute risks may be higher than would follow from a direct use of the composite sample data.

Therefore, the FAO/WHO Geneva Consultation recommended to include a *variability factor* (v) in the non-probabilistic calculation of an international estimate of short-term intake (*IESTI*) (FAO/WHO 1997). The *IESTI* has been adopted by the Joint Meeting of FAO and WHO experts on Pesticide Residues in food in 1999, and was modified in 2000 to reflect that the supply for actual consumption on a given day is likely to be derived from a single lot (JMPR 1999, 2000). In both the original and the modified definition, the variability factor is used in a similar way. The basic idea is that the residue concentration for the first unit eaten is multiplied by v, whereas this factor is not applied for any remaining part of the daily consumption.

In the original presentation v was meant to reflect "*the ratio of a highest level of residue in the individual commodity unit to the corresponding residue level seen in the composite sample*" (FAO/WHO 1997). It was not clearly stated what was meant with "*a highest level*". Should this be the maximum level found or should it be a high percentile, e.g. p95 or p97.5? In practical terms this did not matter too much, because little data were available. Therefore the FAO/WHO Consultation recommended to take *initial* values of v equal to "*the number of units in the composite sample as given in Codex sampling protocols*". This will provide a conservative estimate of the residue concentration in the first unit, based on the assumption that all of the residues present in the composite sample are present in this single unit. If Codex sampling protocols are used, then the number of units per composite sample is 5 for large crops (unit weights > 250 g) and 10 for medium crops (unit weights 25-250 g). For small crops (< 25 g) a variability factor v = 1 was recommended. More recently, it has been proposed to replace the default value 10 with 7. For commodities which are processed in large batches, e.g. juicing, marmalade/jam, sauce/puree, a variability factor v = 1 is proposed. To summarise:

unit weight, *wu*	FAO/WHO default variability factor, *v*
< 25 g	1
25 –250 g	7
> 250 g	5
juicing, marmalade/jam, sauce/puree	1

Table 1: Default variability factors for IESTI calculations.

The Consultation specifically recommended to replace these default values with more realistic values obtained from studies on actually measured units. A working group of the International Conference on Pesticide Residues Variability and Acute Dietary Risk Assessment held in York in 1998 suggested to define v, for samples taken from controlled trials, as the 97.5^{th} percentile of the unit levels divided by the sample mean (Harris et al. 2000), and this is used in the current version of MCRA as the defining relation.

3.2.3.2 Approaches to unit variability in probabilistic modelling

How should variability between units be incorporated in probabilistic modelling of acute risks? In probabilistic modelling we generate consumption amounts and residue concentrations which will be multiplied and summed over products to estimate the intake. However, the residue concentration cm_k will usually be derived from a distribution based on measurements on composite samples. Assume that a batch of product contains N units (N large, for the statistics we assume infinite). The monitoring measurement cm_k is made on a composite sample of nu_k units (for example, nu_k = 5). These units are assumed to be representative of the batch. Unit concentrations cr_{ijk} are to be simulated for one or more units from this batch that will be part of a consumption portion in the Monte Carlo simulation. Basically, there are three possibilities depending on the availability of data:

1. use actual measurement data on individual units;
2. use variability factors or other summary statistics based on measured individual units;
3. use conservative assumptions.

In MCRA only methods under categories 2 and 3 are implemented. The first approach has been pioneered in the context of a large UK survey on pesticides in fruit (Hamey 2000).

In MCRA the following three models, discussed below in more detail, are implemented:

1. **Beta model**, requires knowledge of the number of units in a composite sample, and of the variability between units (realistic or conservative estimates);
2. **Bernoulli model**, requires only knowledge of the number of units in a composite sample (results are always conservative);
3. **Lognormal model**, requires only knowledge of the variability between units (realistic or conservative estimates).

Preferably realistic estimates of unit variability are to be used, either expressed as coefficients of variation *cv* (standard deviation divided by mean) or as variability factors *v* (defined in MCRA as 97.5th percentile divided by mean). However, often such information is not directly available. In such cases it is customary to select high values for the variability factor, either based on collections of variability factors for other compounds/products, or calculated as the theoretical maximum derived from the number of units in a composite sample.

How to translate the concept of conservatism to the probabilistic model? In a non-probabilistic model a higher value of *v* gives a higher *IESTI*, but in a stochastic model a higher variability means more spread around a central value. In general this means that higher values, but also lower values can be generated. In order to retain an overall conservatism it is therefore necessary to replace all simulated values below the monitoring level (cm_k) with cm_k itself.

It is common to use default conservative values, such as the FAO/WHO variability factors in Table 1. However, one should be aware that two entirely different interpretations are possible:

1. The default variability factor may be defined in the same way as a data-based variability factor (*v* = 97.5th percentile/mean). For example, it may be an expert opinion based on seeing many actual data sets from trials, that a certain value v can be used as a conservative value for other situations (see e.g. Table 1 in Harris et al. 2000). Then we might use the beta or the lognormal model, censoring these distributions at cm_k to guarantee conservative behaviour. For the beta model additional information on the number of units in a composite sample is needed.
2. Alternatively, one can revert to the original definition and interpret FAO/WHO variability factors as the number of units in the composite sample ($v = nu_k$). In this case, without other information, the only workable model is the Bernoulli model.

3.2.3.2.1 Beta model for unit variability

With this model MCRA will generate values for individual unmeasured units of a measured composite sample. If cm_k is the concentration measured (or simulated) for the composite sample in monitoring for commodity *k*, then the concentration in any unit can be no larger than $c_{max} = nu_k * cm_k$, where nu_k is the number of units in the composite sample. Under the Beta model simulated unit values are drawn from a bounded distribution on the interval (0, c_{max}). The parameter for unit variability is specified as a coefficient of variation cv_k of the unit values in the composite sample, or as a variability factor.

The standard beta distribution is defined on the interval (0, 1) and is usually characterised by two parameters *a* and *b,* with *a>0 , b>0* (see *e.g.* Mood *et al.* 1974). Alternatively, it can be parameterised by the mean $\mu=a/(a+b)$ and the variance $\sigma^2=ab(a+b+1)^{-1}(a+b)^{-2}$, or, as applied in MCRA, by the mean μ and the squared coefficient of variation $cv^2=ba^{-1}(a+b+1)^{-1}$. Note that the coefficient of variation is the same for the unscaled and the scaled distributions.

For the simulated unit values in each iteration of the program we require an expected value cm_k. This scales down to a mean value $\mu = cm_k/c_{max} = 1/nu_k$ in the (standard) beta distribution. From this value for μ and an externally specified value for cv_k the parameters *a* and *b* of the beta distribution are calculated as:

$$a = b\left(nu_k - 1\right)^{-1}$$

$$b = \frac{\left(nu_k - 1\right)\left(nu_k - 1 - cv_k^2\right)}{nu_k cv_k^2}$$

From the second formula it can be seen that cv_k should not be larger than $\sqrt{nu_k - 1}$ in order to avoid negative values for *b*.

When the unit variability is specified by a variability factor $v_k = \frac{p97.5_k}{cm_k}$ instead of a coefficient of variation cv_k then MCRA applies a bisection algorithm to find a such that the cumulative probability $P[Beta(a,b)] = 0.975$ for $b = a(nu_k - 1)$.
Sampled values from the beta distribution are rescaled by multiplication with c_{max} to unit concentrations cr_{ijk} on the interval $(0,c_{max})$.
In the case that variability has been estimated by a conservative high value, all sampled values lower than cm_k are replaced by cm_k.

In Figure 2 for several values of the coefficient of variation and number of units the beta distribution is shown with estimated parameters a and b. When the parameter for unit variability is high (upper left plot) the ratio of the spikes on the extremes (3:1) represent the 75% probability at $cr_{ijk} = cm_k$ and 25% probability at $cr_{ijk} = c_{max}$. In the upper right plot, the parameter for unit variability is smaller and some unit values in between the two extremes are sampled. The ratio of the spikes is about 5:1, which is according to the number of units in the composite sample. In the lower left plot, variability is low and unit values are sampled around the monitoring residue. In the extreme case, when unit variability is close to zero the monitoring residue itself is sampled and a spike occurs (not shown). The lower right plot show an intermediate situation, moderate to high variability.

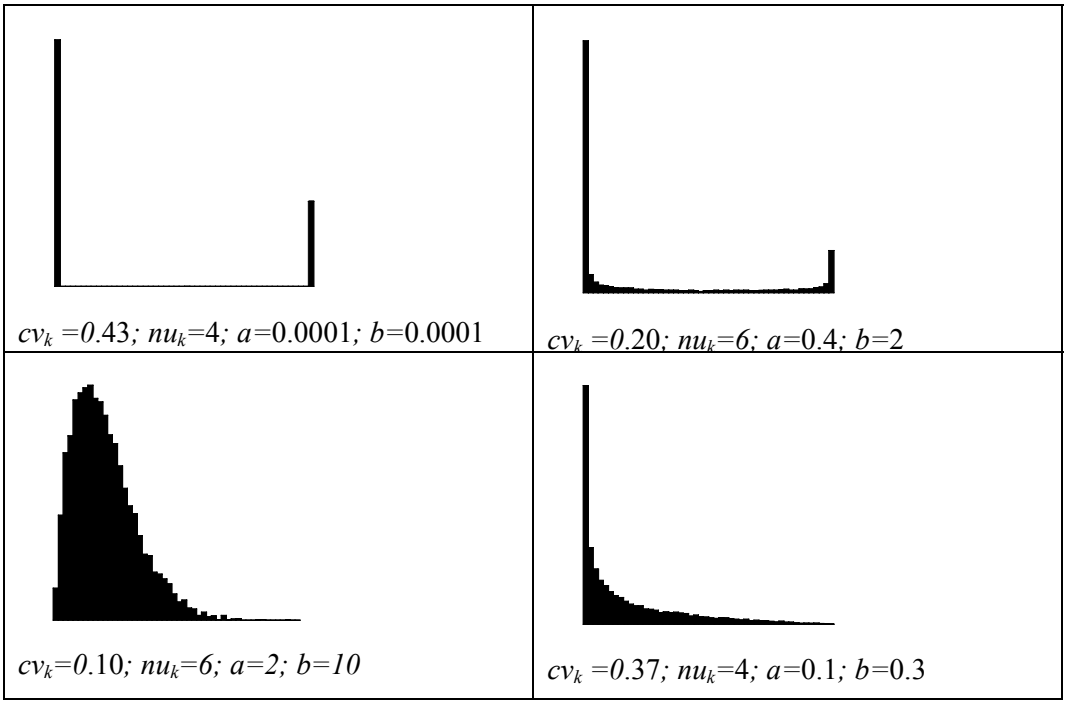


Figure 2: Standard Beta distribution for different values of the coefficient of variation cv_k and number of units nu_k in the composite sample. x axis from 0 to 1.

3.2.3.2.2 Bernoulli model for unit variability
The Bernoulli model is a limiting case of the Beta model, which can be used if no information on unit variability is available, but only the number of units in a composite sample is known. As a worst case approach we may take cv_k as large as possible. When cv_k is equal to the maximum possible value $\sqrt{nu_k - 1}$, the (unstandardised) Beta distribution simplifies to a Bernoulli distribution with probability $(nu_k - 1)/nu_k$ for the value 0 and probability $1/nu_k$ for the value $c_{max} = nu_k * cm_k$..
In MCRA values 0 are actually replaced by cm_k, to keep all values on the conservative side. For example, with $nu_k = 5$, there will be 80% probability at $cr_{ijk} = cm_k$ and 20% probability at $cr_{ijk} = c_{max}$.

3.2.3.2.3 Lognormal model for unit variability

With the Beta and Bernoulli models, MCRA simulates concentrations for units in the composite sample, such that the residue level of an individual unit can never be higher than the monitoring measurement multiplied by the number of units in the composite sample $c_{max} = nu_k * cm_k$.

With the Lognormal model for unit variability MCRA simulates concentrations for new units in the batch from which the composite sample was taken. Effectively the number of units in a batch is very large, so in this case there is no practical upper limit to the residue level that can be present.

The lognormal distribution is considered as an appropriate model for many empirical positive residue level distributions. With the Lognormal model MCRA assumes a lognormal distribution for unit residue concentrations. Let this distribution be characterised by μ and σ, which are the mean and standard deviation of the log-transformed concentrations. The unit log-concentrations are drawn from a normal distribution with mean $\mu = \ln(cm_{ik})$.

Also for the Lognormal model MCRA allows two choices to specify the parameter for the unit variability. The parameter is specified as a coefficient of variation (cv_k) or as a variability factor (v_k). The coefficient of variation cv is turned into the standard deviation σ on the log-transformed scale with:

$$\sigma = \sqrt{\ln(cv^2 + 1)}$$

A variability factor v is converted into the standard deviation σ as follows:

$$v = \frac{p97.5}{mean} = \frac{e^{\mu+1.96\sigma}}{e^{\mu+1/2\sigma^2}} = e^{1.96\sigma-1/2\sigma^2}$$

with μ and σ representing the mean and standard deviation of the log-transformed concentrations. So

$$\ln(v) = 1.96\sigma - 1/2\sigma^2$$

Solving for σ gives: $\sigma^2 - 2*1.96\sigma - 2\log(v) = 0$, with roots for σ according to:

$$\sigma = 1.96 \pm \sqrt{(1.96^2 + 2\log(v))}$$

The smallest positive root is taken as an estimate for σ (see also 3.5.3.2).

In the case that variability has been estimated by a conservative high value, all sampled values lower than cm_k are replaced by cm_k.

3.2.3.3 Estimation of intake values using the concept of unit variability

- For each iteration i in the Monte Carlo simulation, obtain for each commodity k a simulated intake x_{ik} , and a simulated composite sample residue concentration cm_{ik} .
- Calculate the number of unit intakes nux_{ik} in x_{ik} (round upwards) and set weights w_{ikl} equal to unit weight wu_k, except for the last partial intake, which has weight $w_{ikl} = x_{ik} - (nux_{ik} - 1)wu_k$.
- For the Beta or Bernoulli distribution: draw nux_{ik} simulated values κ_{ikl} from a Beta or Bernoulli distribution. Calculate concentration values as $c_{ikl} = \kappa_{ikl} * cm_{k,\,max} = \kappa_{ikl} * cm_k * nu_k$. Sum to obtain the simulated concentration in the consumed portion:

$$cr_{ik} = \sum_{l=1}^{nux_{ik}} w_{ikl} c_{ikl} \Big/ x_{ik}$$

- For the Lognormal distribution: draw nux_{ik} simulated logconcentration values lc_{ikl} from a normal distribution with mean $\mu = \ln(cm_{ik})$ and standard deviation σ. Backtransform and sum to obtain the simulated concentration in the consumed portion:

$$cr_{ik} = \sum_{l=1}^{nux_{ik}} w_{ikl} e^{lc_{ikl}} \Big/ x_{ik}$$

3.2.4 Modelling of missing data and non-detects

Missing data should be indicated by 9999 in the database tables. In principle such values are ignored in the analysis.

Most monitoring measurements of chemical substances are nondetects, i.e. no quantitative measurement is reported. For this reason data are entered in the Concentration table by specifying the total number of measurements made together with the limit of reporting (LOR). We use LOR to mean exactly what the term says: measurements below LOR are not reported, whereas values equal to or higher than LOR are represented by numerical values in the database.

In the analytical and food risk fields analytical limits are often indicated as LOD (limit of detection) or LOQ (limit of quantification). Unfortunately, it is not always clear what is meant with these terms. In any case official recommendations are to always report any available numerical values even if they are below LOD or LOQ limits (IUPAC 1995).

For legal applications of compounds data may be available about the percentage of the crop which receives treatment. When a chemical substance can enter the food chain only via crop treatment, and when the percentage of crop treated is (approximately) known to be $100p_{crop\text{-}treated}$, then this knowledge may be used to infer that $100(1\text{-}p_{crop\text{-}treated})$% of the monitoring measurements should be real zeroes, contributing nothing to pesticide intake, whereas other nondetects in the monitoring data could have any value below the LOR. For $100(p_{non\text{-}detect} + p_{crop\text{-}treated} - 100)$% of the monitoring measurements, 0 and LOR represent best-case and worst-case estimates. A simple way (tier 1 approach) to consider the uncertainty associated with nondetects is to compare intake distributions for these best-case and worst-case situations.

3.3 Comparison of probabilistic with deterministic estimates of acute risk

The IESTI (International Estimated Short-Term Intake) is a deterministic estimate of the short-term intake of a residue on the basis of the assumptions of high daily food consumption per person and highest residues from supervised trials. The IESTI is expressed per kg body weight and has only been defined for single products.

MCRA calculates IESTI for comparison with Monte Carlo percentiles.

Calculations of IESTI (according to FAO 2002) recognise four different case (1, 2a, 2b and 3). In cases 1 to 3 the following definitions are used:

LP: Highest large portion reported, calculated as the 97.5th percentile of the distribution of consumed portions on days with positive consumption of the product (kg food/day)

HR: Highest residue in composite sample, mg/kg

bw: Mean body weight, kg; in MCRA values may be input by the user, or weighted means are calculated over consumers with the number of days on which they consumed the product as weights

U: Unit weight of the edible portion, kg.

v: Variability factor – the factor applied to the composite residue to estimate the residue level in a high-residue unit

MR: Median residue in commodity, mg/kg

Although the FAO Manual refers to supervised trials only, MCRA calculates residue levels (HR or MR) from any residue concentration data set supplied (may also be monitoring data).

Residue levels (HR or MR) may be multiplied with a processing factor on beforehand, in MCRA this depends on the options chosen for processing.

Case 1:
The residue in a composite sample reflects the residue level in meal-sized portion of the commodity (unit weight is below 25 gr).

$$\text{IESTI} = \frac{LP * HR}{bw}$$

Case 2:
The meal sized-portion, such as a single fruit or vegetable unit might have a higher residue than the composite (whole fruit or vegetable unit weight is above 250 gr). Case 2 is further divided into case 2a and 2b.

Case 2a:
Unit edible weight of raw commodity is less than large portion weight.

$$\text{IESTI} = \frac{U * HR * v + (LP - U) * HR}{bw}$$

The formula is based on the assumption that the first unit contains residues at the HR*v level and the next one contains residues at the HR level, which represents the residue in the composite from the same lot as the first one.

Case 2b:
Unit edible weight of raw commodity exceeds large portion weight.

$$\text{IESTI} = \frac{LP * HR * v}{bw}$$

The formula is based on the assumption that there is only one consumed unit and it contains residues at the HR*v level.

Case 3:
For those processed commodities where bulking or blending means that the median represents the likely highest residue.

$$\text{IESTI} = \frac{LP * MR}{bw}$$

When an acute reference dose is available, the calculated IESTI values are also expressed as a percentage of the acute RfD.

IESTI is a deterministic estimate to reflect the unit variability within a composite sample. In the probabilistic approach, unit variability is explicitly modelled and the result is an estimate of the intake distribution (per commodity). These two different approaches handle the same problem, but it is undefined to which Monte Carlo percentile the IESTI value should be compared. In MCRA the user is free to choose a percentage point for this comparison.

A point to note is that IESTI is calculated from positive consumptions on each separate commodity. To allow a fair comparison, the Monte Carlo percentiles are calculated in the same way. Note, however, that in a multi-commodity Monte Carlo analysis, even if one restricts the attention to consumption days only, the percentiles are typically based on consumption data which are partly zero (days with consumption zero for some but not all commodities).

3.4 Binning

Binning is a method to summarise the simulated data (total intake, intake per product, consumption per product, concentration per product) in frequency intervals for further use in deriving the exposure distributions. The alternative would be to store observations for subsequent use, but this would require for moderate simulations already a large amount of storage capacity and an excessive administration.
The mean value of the observations in the first chunk of the simulation (*mean*) is used to define the left limit of the central bin. For values above the mean, 1100 bins are used for storage. The upper limits of the upper bins are defined as 1 % higher than the lower limit. So, for upper bin *i* the upper limit is calculated as *mean* x 1.01^{i}. For values below the mean also 1100 bins are defined with lower limits defined by *mean* x 1.01^{-i}. After the process of binning is completed, the quantile value of a specific percentile is determined by linear interpolation between the bin limits. These 2200 bins together provide efficient storage for numbers spanning more than 9 decades ($1.01^{2200}=3.2\text{x}10^{9}$), which should be amply sufficient for most practical problems.
To get accurate results, it is rather important that the mean value in the first chunk represents, approximately, the true mean of the sampled distribution. Therefore, chunk size (defined as the total number of simulations divided by the number of chunks) should not be chosen too small. During the simulation, the maximum of the sampled observations in each chunk is calculated. When this value is higher than the upper limit of the last bin, representing a potential maximum, this bin limit is replaced by the new maximum, and a warning is issued. When the mean value is missing, e.g. due to zero intakes, the program resorts to an average *mean* value, e.g. the average of the mean values of commodities with nonzero intakes. Also in this case a warning is given.

3.5 How to deal with limited information

In the probabilistic model, a distribution of food consumption data as well as a distribution of residue data are used. For both components of the model, a choice can be made between a non-parametric (see 3.2.1.1) or a parametric (see 3.2.1.2) approach. In a *parametric approach* the data are modelled with an appropriate distributional form (e.g. lognormal with parameters σ and μ). In a *non-parametric approach* the empirical distribution is used to sample from directly. Obviously, the latter approach requires more data to obtain a satisfying representation of the full distribution. Therefore, parametric modelling becomes important in data-scarce situations (see 3.5.1).
Occasionally, limited information emerge not as a consequence of the amount of data but how they are presented: data are reported using e.g. the mean and variance (see 3.5.3) or data are summarised as counts of observations falling into a series of classes (see 3.5.2). It is evident that a parametric approach is the only way out and that the parameters of the lognormal distribution should be inferred using the available data.
If for some commodities there are far less data than for others, it may be sensible to consider *pooling* procedures for means and or variances of the concentration distributions (see 3.5.4).
If the amount of data is limited, this may lead to a relatively large *sampling uncertainty*. Bootstrap procedures may be used to assess the magnitude of this uncertainty (see 3.5.5).

3.5.1 The choice between a parametric and non-parametric approach

How many residue data are required for a sensible calculation of upper-tail percentiles in the exposure distribution based on a non-parametric approach? The rule of thumb can be used that the chosen percentile should be contained directly in the data. For example, at least 20 measurements are needed to estimate the 95th percentile and at least 100 measurements to estimate the 99th percentile. More generally, the number of measurements per food commodity (n) should at least equal $1/(1\text{-}p\%/100)$ to allow a rough empirical estimate of the p^{th} percentile of the residue concentration distribution to be made. Of course, the risk assessment is only coarse with this minimum amount of data and larger sample sizes per food commodity are certainly worthwhile.
In situations where the number of measurements becomes a problem, an appropriate risk analysis should be based on further modelling. Essentially, the lack of data is compensated by *a priori* assumptions. Assuming a simple distributional form for the residue data, the number of measurements can be smaller in principle (at least 10, say). However, non-detect measurements provide no information about variability, and therefore we should now count the number of positive

measurements. Figure 3 shows which approach could be best used depending on the total number of measurements and the number of non-zero measurements. In principle, such a choice could be made separately for each food commodity.

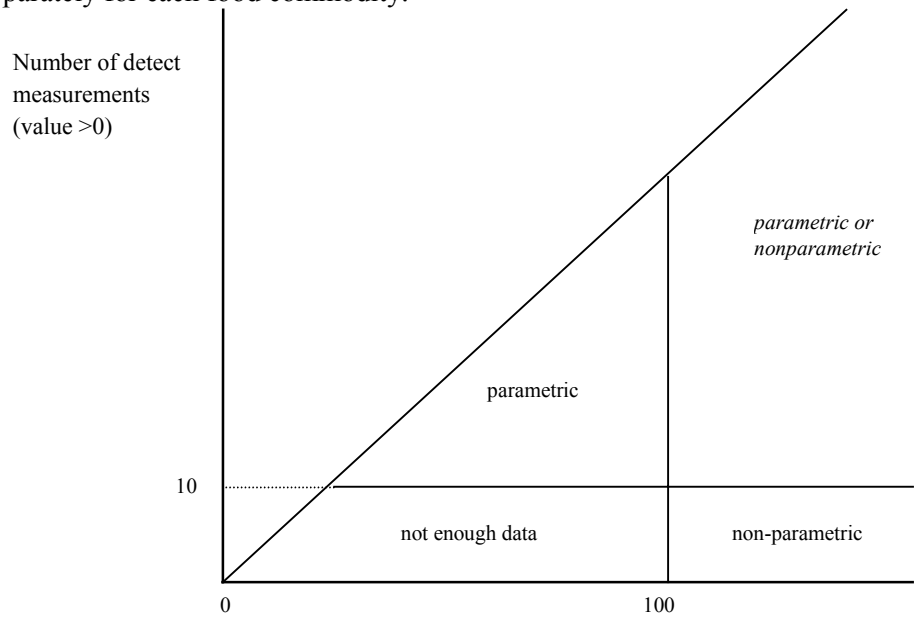


Figure 3: Use of non-parametric or parametric modelling for estimating 99 % exposure percentile in relation to sample size and number of positive measurements.

3.5.2 Estimation based on histogram data

In EU reporting, residue data are sometimes reported in a tabulated (histogram) form: data are expressed as counts of observations falling into a series of groups. The observed counts are $n_1 \ldots n_c$, which fall into c classes with limits $c_1 \ldots c_c$. The number n_1 is the number of positive (*detect*) samples, which are nevertheless below the LOR (= c_1); n_2 is the number of positive samples that fall in between limits c_1 and c_2,…,; n_c is the number of samples that fall in between limits c_{c-1} and c_c.
For histogram data, parameters μ and σ of the lognormal distribution can be obtained by fitting a normal distribution to a set of observations $n_1 \ldots n_c$. In an iterative way, expected counts for a standardised normal variable are calculated using the log-transformed group limits. Each round, parameters are updated until the process converges.

3.5.3 Estimation based on summary data

Occasionally, data are reported in a very condensed form. Summary statistics like the mean, quantiles and dispersion measures as the variance or the coefficient of variation are used to describe characteristics of the underlying residue distributions. The reported statistics are calculated using all values (with concentrations below LOR sometimes replaced by ½*LOR), or using positive values (detects) only. In order to use the binomial-lognormal model, summary statistics based on all values must be corrected for the values replacing the concentrations below LOR. For the mean, the correction is straightforward, taking a zero or the midpoint-value (½*LOR). Likewise, the standard deviation or any measure of dispersion is corrected for the sum of squares due to all zero values and taking into account the corrected mean. The median is also corrected, but instead of correcting the value itself, a corrected quantile z_q is calculated corresponding to q, the lower fraction and z_q satisfying:

$$z_q = \Phi^{-1}\{q\} = \Phi^{-1}\{(\tfrac{1}{2}N - n_0)/(N - n_0)\}$$

with $\Phi(.)$, the cumulative probability function of the standard normal distribution, N, the total number of samples and n_0, the number of zero's.

The maximum is the largest order statistic. Its expected value can be approximated by taking the appropriate population quantile, especially in large samples. Here, the problem is the other way around: the population quantile corresponding to the largest value given the sample size is to be estimated. For sufficiently large N an approximation to $E(q_{max})$ is provided by the value of z_q satisfying $\Phi(z_q) = N/(N+1)$. Blom (1958) and Harter (1961) made the following suggestions for smaller sample sizes:

$$z_q = \Phi^{-1}\{(N - \alpha)/(N - 2\alpha + 1)\}$$

With $\alpha = .315065 + .057974u - .009776u^2$ and $u = \log_{10}N$. Over a wide range of N α approximates the value 3/8. This empirical formula is a very accurate approximation to the exact value of $E(q_{max})$ and is used to estimated appropriate population quantiles for q_{max}. (David, 1970; Pearson and Hartley, 1972 ; Blom, 1958; Harter, 1961).

Three situations can be distinguished:

1) the reported information is insufficient to estimate both μ and σ, or
2) the reported statistics are sufficient to extract μ and σ, or
3) the information is redundant so various estimates for μ and σ are available.

Here, we first consider approaches for situation 2. Situation 1 requires additional information: a solution might be to use the information on comparable product-residue combinations to assess the necessary estimates. Situation 3, basically, is a pooling problem how to weigh and combine estimates that originate from different statistics.

3.5.3.1 Moments and other characteristics

A positive random variable X is said to be lognormally distributed with parameters μ and σ^2 if $Y = \ln X$ is normally distributed with mean μ and variance σ^2. The probability density function of X is:

$$f(x) = 1/(\sqrt{2\pi}\sigma x)\ \exp(-(\ln x - \mu)^2/2\sigma^2).$$

The corresponding normal distribution for Y is denoted by $N(\mu, \sigma^2)$.

Estimation of μ and σ using summary statistics is based on equations and characteristics derived from the moment generating function of the lognormal distribution. Required parameters are estimated by solving the formula's of the first two moments for μ and σ.
The following characteristics for variable X derived from the moment generating function are given:

mean:	$\exp(\mu + 1/2\sigma^2)$	(1)
variance:	$\exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$	(2)
mode:	$\exp(\mu - \sigma^2)$	(3)
quantile (q_q):	$\exp(\mu + z_q\sigma)$,	(4)
vc:	$\sqrt{(\exp(\sigma^2) - 1)}$	(5)

with vc the coefficient of variation, q a given lower fraction and z_q the corresponding standard normal deviate. The 50^{th} quantile, the median, is a special case with $z_q = 0$. The geometric mean of X is equal to the median.

3.5.3.2 Estimation

Approach 1: estimation based on two quantiles, $q_{q1} \neq q_{q2}$.

Using (4) gives:

$\sigma = \log(q_{q1}/q_{q2}) / (z_{q1} - z_{q2})$. Substituting σ yields μ.

Approach 2: estimation based on a quantile and the mean.

Solving for σ using (1) and (4) gives:

$\sigma^2 - 2z_q\sigma - 2\log(\text{mean}/q_q) = 0$, with roots for σ according to:

$$z_q \pm \sqrt{(z_q^2 + 2\log(\text{mean}/q_q))} \qquad (6)$$

For moderate to small sample sizes the estimation of σ fails because the discriminant is negative, i.c. the argument of the square root function. Empirical simulations show that a negative discriminant happens more often for small sample sizes and for estimation based on extreme quantiles like the maximum. Figure 4 shows the empirical relation between the sample size and the fraction of failures for estimation methods involving the mean with respectively, the maximum and median. For the maximum, failures occur already at sample sizes n = 30 to 40, for the median n = 15 to 20. Negative discriminants occur when estimation is based on empirical (sampled) values instead of theoretical (calculated) values assuming a normal underlying distribution. The amount of failures for small sample sizes is in accordance with large sample theory. When the maximum is involved and estimation fails, an estimate of σ is assessed by equating the discriminant to zero. Empirical results show that this works out very well for sample size n > 4, although σ is slightly biased upwards being a conservative estimate. In case of the median no solution to this problem is available so the estimate of σ is set to a missing value.

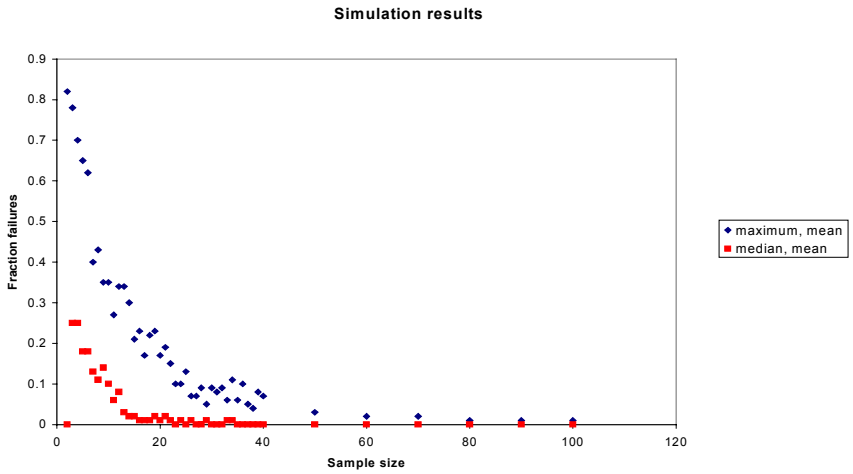


Figure 4: Simulated fraction of failures versus sample size for estimation of σ based on the mean and respectively the maximum and median

In general, for n large enough, say n > 40, σ has two roots. Usually, the mean is larger than the median. Then, σ is estimated with:

$z_q + \sqrt{(z_q^2 + 2\log(\text{mean/median}))}$ with condition $\sigma > 2z_q$..

In case of the mean and maximum σ is estimated with:

$z_q - \sqrt{(z_q^2 + 2\log(\text{mean/max}))}$ with condition $\sigma < 2z_q$..

Note, that max is always greater than the mean. Here, the smallest root is taken as an estimate because empirical results show that the largest root yields unlikely high measures of dispersion and therefore should be rejected.

Approach 3: estimation based on mean and variance or coefficient of variation.

The coefficient of variation, vc = $\sqrt{}$(variance)/mean. Using (5), parameter σ is estimated with:

$$\sqrt{(\log(vc^2 + 1))}$$

and μ is estimated solving (1).

Approach 4: estimation based on a quantile and coefficient of variation[1].

For estimation of σ, see approach 3. Using (4), parameter μ is estimated with:

$$\log(\text{quantile}) - z_q\sigma$$

For the median, estimation of μ simplifies to:

$$\log(\text{median})$$

3.5.4 Grouping of products, pooling of means and variances

When data are limited, it may be advantageous to apply the parametric approach for modelling of the positive concentrations. In MCRA the positive concentrations are modelled as lognormal with parameters μ and σ^2, representing mean and variance of the natural logarithm of the concentrations. However, estimation is often hampered because data on residues in specific food commodities are sparse or even missing. In those cases, grouping of products into product groups enlarges the number of measurements per group and may give sufficient data to base estimates upon. We must assume that residue distributions are the same for the grouped products. A second related question is the reliability of estimates, based on a few number of degrees of freedom. The following procedure is designed to cope with the above problems.

1. **Pooling variances within product groups.** For each product the variance σ^2 and mean μ is estimated. Then, products are assigned to product groups which are composed of related products, e.g. a productgroup containing sorts of cabbages or a group containing all kind of berries. Products where agricultural use is allowed are remained separate from products where agricultural use is not allowed. The homogeneity of variances in the different product groups is assessed using Bartlett's test (Snedecor & Cochran, 1980). The test statistic determines whether variances within a group are to be pooled automatically ($p > 0.05$) or not ($p \leq 0.05$).
2. **Pooling means within product groups.** After pooling the variances, an overall test for differences of means within each group is performed, based on analysis of variance. Means within groups are pooled automatically if the probability $p > 0.05$.
3. **Using overall variance if there are < 10 degrees of freedom.** Estimates of variances based on less than 10 df are considered not very reliable. Therefore, variances based on < 10 df are compared to the overall variance (pooled over all products except the tested product itself, i.c. corrected) and tested for equality. Variances are replaced by the overall variance (uncorrected) whenever the hypothesis of equality of variances is not rejected; if rejected, the original variances are maintained.

[1] Not yet implemented

For a parametric risk assessment all variances and means must be present. This requirement implies that very often rearrangement of products into (sub)groups preceeds the actual simulation of the intake distribution.

To summarise, actions are:
- calculate variances and means for each product
- classify products into groups
- test homogeneity of variances and equality of means within groups of products. Results are: not significant ($p > 0.05$) or significant ($p \leq 0.05$).
- take products(-groups) with $df < 10$
- compare variance with overall variance (corrected). Replace variance with overall variance (uncorrected) for non-significant test results.

3.5.5 Assessing the uncertainty of risk assessments by bootstrapping data sets

In probabilistic risk assessment of dietary intake we use distributions which describe the *variability* in consumption within a given population of consumers and the *variability* of the occurrence and level of chemical residues on the consumed commodities. However, these calculations do not consider the amount of *uncertainty* that is due to the limited size of the underlying datasets. Typically, in a large number of simulations very many different combinations of consumption and residue concentrations are made. This leads to a smooth distribution of simulated intakes, and the impression of a very precise estimation of exposure percentiles or other quantities of interest. It is essential to realise that the accuracy of the inference depends on the accuracy of the basic data.

A computer-based instrument to assess the reliability of outcomes is the *bootstrap* (Efron 1979, Efron & Tibshirani 1993). In its most simple, non-parametric form, the bootstrap algorithm resamples a dataset of n observations to obtain a *bootstrap sample* of again n observations (sampling with replacement, that is: each observation has a probability of $1/n$ to be selected at any position in the new bootstrap sample). By repeating this process B times, one can obtain B bootstrap samples, which may be considered as alternative data sets that might have been obtained during sampling from the population of interest. Any statistic that can be calculated from the original dataset (e.g. the mean, the standard deviation, the 95th percentile, etc.) can also be calculated from each of the B bootstrap samples. This generates a *bootstrap distribution* for the statistic under consideration. The bootstrap distribution characterises the uncertainty of the inference due to the sampling uncertainty of the original dataset: it shows which statistics could have been obtained if random sampling from the population would have generated another sample than the one actually observed.

In Monte Carlo modelling of acute risks two datasets are combined: consumption data and residue concentration data. It makes sense to apply bootstrapping to both datasets separately, in order to characterise the uncertainty in the final exposure. In MCRA the bootstrap algorithm (when selected) is applied to:
1. the multivariate consumption patterns and associated body weights: actually the data set of consumer identifiers is bootstrapped, and all consumer information (consumption patterns for all consumption days, body weight, age) is coupled to the selected consumer identifiers.
2. the univariate residue concentration data sets: these are bootstrapped independently for all commodities. In principle, the bootstrap algorithm is applied to the dataset consisting of both nondetects and positive values; in practice, for a dataset with n_0 nondetects and n_1 positive values, the number of positive values in a bootstrap sample is obtained as a draw from a binomial distribution with parameter $n_1/(n_0 + n_1)$ and binomial total $n_0 + n_1$. Then, this number of values is selected randomly from the set of n_1 positive values.

In MCRA the resulting bootstrap distribution of percentiles of the exposure distribution is summarised by specifying empirical 2.5th , 25th, 75th and 97.5th percentiles. The outer percentiles constitute a central 95% confidence interval for the variability percentiles. However, for this it is necessary that the number of bootstrap samples B is high enough. The number of bootstrap samples should be

chosen depending on the confidence level wanted for the uncertainty interval. Typically 500-2000 bootstrap sets will be reasonable for a 95 % confidence interval (Efron & Tibshirani 1993, pp. 14-15, 275).

The same bootstrap algorithm can also be applied to deterministic estimates which are calculated from data sets. For example the maximum residue value found in a bootstrap sample will be different, if the actual maximum value in the original dataset has *not* been selected. Also data-based estimates of large portion and average body weight will vary.

3.6 Chronic risk assessment

So far, we discussed a probabilistic model allowing for effects of processing, incorporating unit variability and using information on percent crop treated. Through Monte Carlo simulations an estimate of the probability distribution is generated to assess acute risks by intake of the chemical substance. With the right assumptions, we may use this concept for chronic risk assessment.

3.6.1 Introduction

In dietary risk assessment, usual intake is defined as the long-run average of daily intakes of a dietary component by an individual. From a statistical point of view, assessing the usual intake can be reduced to the problem of estimating the distribution of a random variable y_i that is measured with error. A model for the relationship between the observations y_{ij} and the true random variable of interest y_i is:

$$y_{ij} = y_i + u_{ij}$$

where u_{ij} is an additive measurement error for individual i on day j. For independent, normally distributed y_i and u_{ij}, estimation of the distribution of y_i is straightforward. When observations y_{ij} are non-normal and the measurement error variance is heterogeneous across sampling units, estimation is less simple. Nusser *et al.* (1996) describe a procedure for estimating the percentiles of the distribution of long-run average daily intakes using non-normal dietary intake data. Principally, their method consists of three steps:

1. transforming the daily intake data to approximate normality using a combination of a power function and a grafted polynomial function. The polynomial provides some flexibility against power transformed components that are still deviating from normality,
2. estimating the parameters of the usual intake distribution in the transformed scale, and
3. estimating the percentiles of the distribution of usual intakes in the original scale.

The basic ideas of Nusser *et al*. are suited for dietary components that are consumed on a nearly daily basis, e.g. dioxin in fish, meat or diary products.

3.6.2 Modelling long term daily intake

Thanks to the assumed independence of consumption and residue concentrations (a most reasonable assumption) the modelling of usual intake simplifies to a univariate stochastic model for intakes
Usually, food consumption data are available for individuals on 2 (or more) consecutive days. For each individual, the intake on day j is estimated by multiplying the consumptions with the average value of the residues (detects and nondetects) found on each commodity. We assume an equal number of days for each individual. This is in confirmity with our method of reporting consumption only. As a consequence, days without consumptions do have zero intake.

The model for the usual intake distribution is:

$$y_{ij} = y_i + u_{ij}$$

with y_{ij} the observed intake of individual i on day j, y_i is the unobservable usual intake value for individual i, and u_{ij} is the unobservable measurement error for individual i with $i = 1…n$ on day j with $j = 1…o$. In the normal scale, $y_i \sim N(\mu, \sigma^2_{cons})$, $u_{ij} \sim N(0, \sigma^2_{day})$.

3.6.2.1 Step 1: power transformation and spline function

The observations y_{ij} are transformed close to normality using a power transformation. As indicated by Tukey (1962), the expected value of a normal score $z = (y-\mu)/\sigma$ can be approximated by the U-score:

$$U_{ij} = \Phi^{-1}[(r_l - 3/8)/(N + 1/4)]$$

where r_l is the rank of the ij^{th} observation y_{ij} and N, the total number of observations (no. individuals x no. days). The power γ is estimated by minimising the error sum of squares:

$$\sum_{i=1}^{n} \sum_{j=1}^{o} (u_{(ij)} - \beta_0 - \beta_1 y_{(ij)}^{\gamma})^2$$

over a grid of values of γ, where $U_{(ij)}$ and $y_{(ij)}$ denote the order statistics of U_{ij} and y_{ij}.
The observations are replaced by power transformed observations:

$$z_{ij} = y_{ij}^{\gamma}$$

After a power transformation, some components still deviate from normality. To minimise deviations in the Y-direction an integrated B-spline is fitted to the (U_{ij}, z_{ij}) pairs. The spline function is enforced to be monotone increasing by constraining the parameters to be nonnegative. The knots of the spline function are placed such that the interval lengths between knots are equal with two data points left to the left knot and two right to the right knot. The number of knots is optional, here K = 7 is taken. In the intervals, a cubic spline of order 3 is fitted, outside the joint left and right knot the spline is linear. Observations that are transformed by a power in combination with a spline function are denoted by $z_{spline,ij}$. These values are approximate normally distributed.

3.6.2.2 Step 2: estimation of parameters of the usual intake distribution

The power transformed daily intakes are transformed having zero mean and unit variance:

$$z^{*}_{spline,ij} = (z_{spline,ij} - \hat{\mu}_{spline})/\hat{\sigma}_{spline}$$

Parameters of the standardised usual intake distribution in the normal scale are estimated assuming the following model:

$$z^{*}_{spline,ij} = z_{spline,\,i} + u_{ij}$$

with variance components σ^2_{cons} estimating the variability between consumers, and σ^2_{day}, estimating the day to day variability within consumers. The variance components are estimated using standard statistical methods (ANOVA). Their sum is close to 1 because the transformed data (indicated by the asterisk) have mean 0 and variance 1. Normal equivalent deviates of the usual intake distribution (mean 0 and variance σ^2_{cons}) are calculated using:

$$q_{usual} = \hat{\sigma}_{between}\Phi^{-1}(p/100)$$

with p a percentage and $\hat{\sigma}_{between} = \hat{\sigma}_{cons}$.

3.6.2.3 Step 3: backtransformation and estimation of usual intake

Percentiles in the original scale are estimated by a linear interpolation using the $(U_{ij}, z^*_{spline,ij})$ pairs: q_{usual} specifies the values for which interpolated z^*-values are required. The interpolated standardised values, say $z^*_{usual,\ spline}$, are transformed to the original scale by the inverse of the power and correcting for the variance and the mean of the original variable:

$$z_{usual,\ spline} = (z^*_{usual,\ spline} * \hat{\sigma}_{\ spline} + \hat{\mu}_{\ spline})^{1/\gamma}$$

4 Appendix

4.1 Procedures in the MCRA program

A procedure library MCRA.LIB is attached to the program MCRA. This library contains a collection of procedures which are used during a MCRA session. In the following sections, the procedures are arranged into blocks and shortly described.
LIBDATE: contains date that mcralib is created.

4.1.1 Loading input data

The MCRA User Manual (de Boer *et al*., 2004) describes the format of the database tables needed to perform a risk assessment.

FINDKOMMA: searches position of komma in string.
DBWALDOIMPORT: new version of DBIMPORT. Uses suspend (executable) rather than pass (odbc.dll: bug on the internet with ODBC-driver)
MDBSQL: defines query which is used to select data from multiple tables in database.
PRODLABREAD: reads commodity-labels and levels (PNRLAB, PNRLEV) from table **Products**. Reads indicator (ALLOWED) if residue is allowed or not on a commodity. Reads UNITWEIGHT, EDIBLEPORTION and LARGEPROTION. Prints number of commodities (NPNR).
CMPLABREAD: reads residue-label (SNRLAB) from table **Compounds**. Prints residue code (SNRLEV) and label. Reads ARFD and ADI.
INDIVIDUALSREAD: reads consumer characteristics (PERSNR, PLEEF, PGEWI, SEX) from table **Individuals**. Calculates the total number of consumers contained in the database (PERSTAL) and the minimum and maximum age and weight.
MDBCONSUDATREAD: reads consumption data and processing codes from tables **Foodconsumption, Foodconversionmodel, Foods, Products, Individuals.** Forms a subset containing consumption-days only and prints a warning message. Reads HCPNR, RESP, CONSUM, PROCCODE, PERSNR, HLAB, LEVDAG. Prints number of consumed products.
SDATREAD: reads residue summary data from tables **Summarydata, Products, Compounds**. Calculates parameters μ and σ of the lognormal distribution. Calculates mean residues, fraction of positive values (detects) and number of zero residues (nondetects). Reads LOR.
HDATREAD: reads histogram data on residues from **Histodata, Products, Compounds**. Calculates parameters μ and σ of the lognormal distribution. Calculates mean residues, fraction of positive values (detects) and number of zero residues (nondetects). Reads LOR.
FDATREAD: reads residue concentration data from **Concentrations, Products, Compounds, Country**. Calculates mean residues, fraction of positive values (detects) and number of zero residues (nondetects). Calculates parameters μ and σ of the lognormal distribution. Reads LOR. Calculates PLOR, WORSTCASEDETECTION. Replaces missing LORs by substitute values.
WCDATREAD: reads worstcase values from table **Products**, **Agriculturalworstcase** (Systemfile). Print a warning for products for which worstcase values are used.
WCREPLACE: places worstcase values in WCPOINTER (vmx, maxx, median, Mean).

4.1.2 (Pre-)processing of datastructures

In this block, datastructures are pre-processed. In general, pre-processing is needed whenever the user specifies restrictions or wishes to explore the effect of e.g. processing or unit variability. During pre-processing, warnings may be generated when tables are not according to the requested format or when structures are incompatible due to internal errors or redundant, missing and/or unknown codes, values and/or labels.
NPNRPRO: calculates new number of commodities or number of commodities/processing type combinations.
VFREAD: reads variability factors and number of units in composite sample from table **Products, VariabilityProd, VariabilityCompProd** (exceptions dependent on Compound). Processes missing

values for unit weights and variability factors and implements worstcasedetection. Reads UNITCOMP, VARFACTOR, UNITVARMODEL and PARACVORV. Contains ESTCVBETA.
ESTCVBETA: calculates a coefficient of variation when a variability factor is provided for the beta distribution. Calculations are based on bisection algorithm.
PFLABREAD: reads processing codes, labels and distibution from table **Processing.** Reads variability factors and number of units in composite sample form tables **Products, Compounds, VariabilityProcCompProd** (exceptions dependent on compound and processing type). Makes new labels which are combinations of commodity and type of processing. Checks if codes for consumed processed commodities are supplied and prints a warning message if not. Calculates the total number of commodities and processing type combinations (NPNR) and replaces the old value of NPNR (total number of commodities) by the new value. Forms new variates for unit weights and variability factors by expanding the old structures according to the number of times each commodity is processed. Replaces unit weights and variability factors of processing types 9, 11 and 13 by default values 9999 and 1, respectively. Contains EXCVFREAD.
EXCVFREAD: Processes missing values for unit weights and variability factors and implements worstcasedetection. Reads UNITCOMP, VARFACTOR, UNITVARMODEL and PARACVORV. Contains ESTCVBETA (see also VFREAD)
%CONSUDATREAD: calculates consumption data matrix for unprocessed or processed commodities. Checks if all labels for consumed commodities are present and prints a warning if unknown commodities are present. Creates variate with respondent and daynumbers for option Consumers only. Note that the consumption data matrix contains all available days. Applies age restrictions and performs pre-processing for a printed summary of the data. Prints text structures to temporary file, reads variates from temporary file (temp.tmp).
AGRICUSEREAD: reads percentage crop treated from table **Products, Compounds, Country, Agriculturaluse**.
PFDATREAD: reads available information on processing factors (f_k) from tables **Processingfactor, Products, Compounds** and the type of transformation for distribution based factors. If $f_{k,upp}$ and $f_{k,nom}$ are both missing, a value 1 is inserted; if $f_{k,upp}$ is missing, it is replaced by $f_{k,nom}$ and vice versa; if $f_{k,nom} > f_{k,upp}$ both values are interchanged. For fixed processing factors, $f_k = f_{k,upp}$ for commodities on which processing information is available. Otherwise a default value 1 is inserted. For distribution based factors, means (= log or logit transformed $f_{k,nom}$) and variances (based on $f_{k,upp}$ and $f_{k,nom}$) are calculated. A warning is printed when distribution based factors f_k cannot be sampled because $f_{k,nom}$ is missing. For those commodities fixed values are taken instead. Read PFVAR, PFMEAN, LOGNORMAL.
PRPFLAB: generates print information, e.g. labels for those commodities that are processed in two or more ways.
CHCKS: The value of the upper quantile of the intake distribution needed for the summary report of the upper tail is specified by the user. This value may conflict with the chunksize (S) and simulation size (N).The following rules are used: a constant Q_{max} is defined as S/N*100*2. If the user supplied value is smaller or equal than 100 – Q_{max}, then the current value is replaced by 100 – Q_{max}, and the percentage of the upper tail equals Q_{max}. This rule is applied when the user value is set too low (upper tail is too large). On the other hand, when the supplied upper quantile is set too high compared to the total simulation size, the upper quantile is reset to the default value 99.0%, which is usually sufficient.

4.1.3 Estimation of parameters of the lognormal

For a parametric simulation all parameters must be present. A pooling procedure is applied to provide estimates for μ and σ.
NVHOMOGE: new version of VHOMOGENEITY. Tests homogeneity of variance.
NMHOMOGE: tests homogeneity of means and performs automatically pooling for p >0.05.
POOLING: pools variances and means automatically
TABPOOLING: prints a summary of the data after the pooling procedure (number of detects, nondetects, fraction of detects, pooled parameters μ and σ of the lognormal distribution, the original parameters on logscale before pooling, number of degrees of freedom of sigma after pooling,

productgroups e.g. groups of commodities arranged on common characteristics in combination with allowance of the use of a residue on a commodity).
NOPOOLING: prints a summary of the data (number of detects, nondetects, fraction of detects, parameters μ and σ of the lognormal distribution). For all commodities, parameters μ and σ need to be present because parametric modelling is set without pooling. When some variances are missing, the job is abandoned and a warning message is printed.

4.1.4 Simulation of exposure values

In this part, the exposure distribution is generated. The simulation is performed in chunks specified by the user. In each cycle, relevant output is collected and stored for later use.
GETCONSUMPTION: simulates consumption matrix e.g. selects randomly consumers for a specified day or selects randomly consumers irrespective of day. Samples consumption days-only. Calculates total consumption of each commodity and number of consumption occasions.
PFSIMU: calculates matrix with fixed processing factors or factors based on a normal distribution: for each consumption occasion a processing factor is simulated. Backtransforms values according to applied transformation, e.g. logarithm or logit.
_UNITINTAKES: simulates a residue matrix based on empirical data. Residues are simulated for each consumption. When variability is incorporated in the model, VMAX times a new residue matrix is simulated using the sampled value for a consumption and multiplied with consumer unit portions. If option use variability is *no,* VMAX is set to 1. Prints a message about variability factors. For processed commodities, an expanded matrix is simulated with the number of columns equal to the number of combinations of commodities and processing types. Missing values are replaced by LORs. Residues are multiplied with processing factors. Calculates the total sum of the processing factors and the total number of consumption occasions in order to calculate an mean processing factor. The intake is calculated and the total number of positive residues. Contains VARFAC, LORREPLACE.
P_SIMU: simulates a residue matrix based on parametric modelling. Residues are simulated for each consumption. When variability is incorporated in the model, VMAX times a new residue matrix is simulated using the sampled value for a consumption and multiplied with consumer unit portions. If option use variability is *no,* VMAX is set to 1. Prints a message about variability factors. For processed commodities an expanded matrix is simulated with the number of columns equal to the number of combinations of commodities and processing types. Missing values are replaced by LORs. Residues are multiplied with processing factors. Calculates the total sum of the processing factors and the total number of consumption occasions in order to calculate an mean processing factor. The intake is calculated and the total number of positive residues. Contains VARFAC, LORREPLACE.
VARFAC: calculates the number of units in a consumption and standard deviation based on variability factors. Calculates the maximum number of units (VMAX) of all consumptions irrespective of commodity. Generates print information about the commodity with the maximum number of units found.
LORREPLACE: replaces missing values by LORs. All values are replaced or replacement is based on the percent crop treated. In the latter case, the sum of the percentage of non-zero's (detects) and number of LORs (replaced missing values) only approximately equals the percent crop treated because a randomisation step is involved in assigning LORs to zeros.
RESICALC: generates summary statistics for output.
T4ACALC: performs data processing in each cycle to generate the upper quantile of the intake distribution and consumer characteristics of the top 10 intake. Simulation results of two successive cycles are collected in new structures with two times the length of a chunksize. Then, calculations are performed and various data structures with double length needed to produce output are sorted. The intake results needed to summarise the upper quantile of the intake distribution are saved in structures with the same length as a chunk. In the next cycle, these sorted results and new simulation results are collected again and all calculations are repeated. Note that the process of simulating in cycles restricts the value of the upper quantile. Specifying a too large upper tail may supersede the user supplied value. See also procedure CHCKS.
T4BCALC: performs calculations to summarise the total intake distribution.

4.1.5 IESTI estimation

IESTI estimation.
IE_PFSIMU: applies processing on residues (maximum and median).
UV_IESTI: reads unit weight and large portion form table **Products, VariabilityProd** (for option own variability factors, otherwise factors according to WHO). Estimates IESTI based on all consumption and residue data in case of unit variability.
BIN1COUNT: sets limits for bins based on the mean value in the first chunk. Performs binning.
BIN2COUNT: performs binning based on limits of the first chunk.
BIN3COUNT: calculates Monte Carlo percentiles for specified percentile.
TAB8PRINT: calculates IESTI percentiles and prints information to file.

4.1.6 Chronic risk assessment

Chronic risk assessment and bootstrapping.
GAMMA: performs power transformation on intake distribution.
VCREML: estimates variance components.
%INTERPOLATE: interpolates backtransformed chronic percentiles according to Nusser.
NUSINTAKE: estimates daily intakes of consumers based on consumption and mean residue values.
NUSSER: estimates chronic exposure based on power or logtransformed intakes using a grafted polynomial. Long term exposure is estimated for a specified percentage of nondetects. Prints percentiles and cumulative percentiles for a specified value (years). Contains GAMMA, VCREML, %INTERPOLATE, HTMHEAD, HTMCODE, HTMGEN, HTMBUT (nusserdiag.htm, percentiles.dat, percentiles.htm).
NUSCONS: stores bootstrapped consumptions and coupled weights in matrix.
NUSRES: calculates mean residue values of bootstrapped residuals and performs processing.
NUSBOOTSTRAP: estimates chronic exposure (see NUSSER) and bootstrappercentiles.
NUSOUT: prints a summary of chronic risk exposure and results of bootstrapping.

4.1.7 Generating output

Procedures that generate output and graphics.
TCO: warning message consumption days only.
TAB1PRINT: prints a summary of the data used for simulating consumptions and residues. Mean consumptions are averaged after day and/or age restrictions. Printed output is on commodity, average consumption for all consumers and consumers only, number of consumer occasions, the average residue (corrected for processing and after missing values have been replaced by the LOR), the number of non-zero residues and the total number of samples (non-zero and zero residues). The same information is printed for commodities which are processed in more than one way.
TAB2PRINT: prints a summary of the simulation results. Printed output, see TAB1PRINT. Three columns are added: the first describes the difference (%) compared to the average consumption of the data and the second the difference (%) compared to the average residue of the data, the last gives the average of the processing factors per commodity corrected for consumption ratio's. This table is used to compare the simulation results with the summarised data. Large discrepancies between both tables indicate that simulation results are variable.
TAB3PRINT: prints percentiles, the maximum and average intake. Contains HTMHEAD, HTMCODE, HTMGEN, HTMBUT (percentiles.htm, percentiles.dat).
T4APRINT: prints characteristics of the intake distribution per commodity of the specified upper quantile of the total intake distribution. Printed output is relative contribution of each commodity to the total intake distribution; mean, median, 2.5^{th} and 97.5^{th} percentile of the distribution of each commodity and the percentage of zero's. The same information is printed for commodities which are processed. Contains HTMHEAD, HTMCODE, HTMGEN, HTMBUT, HTM1MOUSE, HTM2MOUSE (averconccomres.htm, averconccomres.dat, uppersens.htm), _BIN3COUNT.
_BIN3COUNT: calculates 2.5^{th}, 50^{th} and 97.5^{th} percentiles for table 4a and 4b.
T4BPRINT: prints characteristics of the total intake distribution per commodity. Printed output, see T4APRINT. Contains HTMHEAD, HTMCODE, HTMGEN, HTMBUT, HTM1MOUSE, HTM2MOUSE (totalsens.htm).

TAB5PRINT: prints the intake per commodity of the 10 consumers with the highest total intake and bodyweight and age. The same information is printed for commodities which are processed in two or more ways.
TAB6PRINT: prints the consumption per commodity of the 10 consumers with the highest total intake. The same information is printed for commodities which are processed in two or more ways.
TAB7PRINT: prints residue levels per commodity of the 10 consumers with the highest total intake. The same information is printed for commodities which are processed in two or more ways.
PLTOTDISTR: plots a graph of the total distribution of positive intakes. Contains HTMHEAD, HTMCODE, HTMGEN, HTMBUT (totaldistr.htm, totaldistr.dat).
PLUPDISTR: plots a graph of the upper tail of the intake distribution. Contains HTMHEAD, HTMCODE, HTMGEN, HTMBUT (upperdistr.htm, upperdistr.dat).

4.1.8 Generating output for Component One

For communication with a browser, CompOne output is written in ActiveX-code and HTML-script. Supporting procedures are:
WARNING: warning message 'Fatal error occurred, see logfile'
INPSHEET: prints a short summary of the specifications
HTMPRINT: pop-up menu to request output
HTMHEAD: header and definitions HTML-pages
HTMCODE: definitions cabinet file and linkage package ComponentOne ActiveX controls
HTMGEN: definitions chartarea ComponentOne ActiveX controls
HTM1MOUSE: definitions mouse control ComponentOne ActiveX controls
HTM2MOUSE: definitions mouse control checkbox ComponentOne ActiveX controls
HTMBUT: definitions button onclick ComponentOne ActiveX controls
STRIP: removes physical server address from the file specification and transscripts the path to a internet-address.
RIGHTVIEW: prints the right frame of the HTML-form to view output
LEFTVIEW: prints the left frame of the HTML-form to view output
MCRAVIEWOUTPUT:defines frames for viewing output
PROGRESS: progress and cpu time for HTML-page.

4.1.9 Additional files

PARA.DAT: contains definitions used in defining queries. Stored in directory mcra_3.1\proc\.
DBIMPORTWALDO.EXE: executable for communications with ODBC-driver (see also DBIMPORTWALDO). Stored in directory mcra_3.1\proc\.
MCRA.LIB: MCRA procedure library. Stored in directory mcra_3.1\proc\.

References

- Bestfit.(1997). Probability distribution fitting for Windows. Pallisade Corporation, Newfield, NY, USA.
- Blom, G. (1958). Statistical estimates and transformed beta-variables. Wiley, New York.
- Boon, P.E., van der Voet, H & van Klaveren, J.D. (2003). Validation of a probabilistic model of dietary exposure to selected pesticides in Dutch infants, *Food Additives and Contaminants*, 20, Suppl. 1: S36-S49.
- Crossley, S.J. (2000). Joint FAO/WHO Geneva consultation – acute dietary intake methodology. *Food Additives and Contaminants*, 17: 557-562.
- David, H.A. (1970). Order statistics. John Wiley & Sons, New York.
- de Boer, W.J., van der Voet, H., P.E. Boon, G. van Donkersgoed & J.D. van Klaveren (2004). MCRA, a web-based program for Monte Carlo Risk Assessment, Release 3, User Manual. Report March 2004. Biometris and RIKILT, Wageningen University and Research Centre, Wageningen.
- Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, **7**: 1-26.
- Efron, B. & Tibshirani, R.J. (1993). An introduction to the bootstrap. *Chapman & Hall*, New York.
- FAO/WHO (1997). Food consumption and exposure assessment of chemicals. Report of an FAO/WHO Consultation, Geneva, Switzerland. 10-14 February 1997.
- FAO (2002). Submission and evaluation of pesticide residues data for the estimation of maximum residue levels in food and feed. FAO, Rome.
- GenStat (2002). GenStat for Windows. Release 6.1, sixth edition, VSN International Ltd., Oxford.
- Hamey, P.Y. (2000). A practical application of probabilistic modelling in assessment of dietary exposure of fruit consumers to pesticide residues. *Food Additives and Contaminants*, 17: 601-610.
- Harris, C. *et al.* (2000). Summary report of the International Conference on pesticide residues variability and acute dietary risk assessment. *Food Additives and Contaminants*, 17: 481-485.
- Harter, H.L. Expected values of normal order statistics. Biometrika 48: 151-165.
- IUPAC (1995). Nomenclature in evaluation of analytical methods including detection and quantification capabilities (IUPAC Recommendations 1995). *Pure and Applied Chemistry* 67: 1699-1723.
- JMPR (1999, 2000). Reports of the joint FAO/WHO meetings of experts on Pesticide residues in food.
- Kistemaker C., Bouman M. and Hulshof K. F. A. M. (1998). De consumptie van afzonderlijke producten door Nederlandse bevolkingsgroepen - Voedselconsumptiepeiling 1997-1998. Zeist, TNO-Voeding (Report No: 98.812).
- Nusser SM, Carriquiry AL, Dodd KW & WA Fuller (1996). A semi-parametric transformation approach to estimating usual daily intake distributions. JASA 91(436): 1440-1449.
- Pearson, E.S. and Hartley, H.O. Biometrika tables for statisticians (1977). Vol II.
- Shimizu, K., and Crow, E.L. (eds). (1988). Lognormal distributions: theory and applications. Marcel Dekker, INC. New York.
- Snedecor, G.W. & Cochran, W.G. (1980). Statistical Methods (7th edition). Iowa State University Press, Ames, Iowa.
- van der Voet, H., de Boer, W.J. & Keizer, L.C.P. (1999). Statistical instruments for dietary risk assessment concerning acute exposure to residues and contaminants. Report August 1999, Centre for Biometry Wageningen, Wageningen.
- van der Voet, H., de Boer, W.J. & Boon, P. (2001). Modelling exposure to pesticides. Note HVT-2001-03, Centre for Biometry Wageningen, Wageningen.
- van Dooren, M. M. H. , Boeijen, I., van Klaveren, J. D. and van Donkersgoed G. (1995). Conversie van consumeerbare voedingsmiddelen naar primaire agrarische produkten. RIKILT-report. Wageningen, RIKILT-DLO (Report No: 95.17).

- van Klaveren, J.D. (1999). Quality programme for agricultural products. Results residue monitoring in the Netherlands. RIKILT Institute of Food Safety, Wageningen.

Errata (28-04-2004)

page 13, figure 2:

upper left: cv_k *=0.43* should be cv_k *=1.732*
a=0.0001 should be a=0.00005
b=0.0001 should b=0.00015

upper right: cv_k = 0.20 should be cv_k = 1.20

lower left: cv_k = 0.10 should be cv_k = 0.62

lower right: cv_k = 0.37 should be cv_k = 1.46