Statistical instruments for dietary risk assessment concerning acute exposure to residues and contaminants

Hilko van der Voet
Waldo J. de Boer
L.C. Paul Keizer

Report August 1999

Agricultural Research Department (DLO)
Centre for Plant Breeding and Reproduction Research (CPRO-DLO)
Centre for Biometry Wageningen (CBW)
P.O. Box 16, 6700 AA Wageningen, The Netherlands
Telephone: +31 317 47 70 00
Fax: +31 317 41 80 94
E-mail: post@cpro.dlo.nl

Table of contents

Summary

A practical implementation has been made of the probabilistic approach to dietary acute risk assessment. Full electronic databases with food consumption data and residue concentration data were available from the Food Consumption Survey (VCP) and the Quality Program for Agricultural Products (KAP), respectively. A non-parametric Monte Carlo approach, using food consumption and residue data from the databases directly, was chosen as a basic method. The implementation in a general-purpose statistical language (Genstat) gave similar results as an implementation using the decision tool @Risk, but performed much faster.

Even with the large amount of data available in the Netherlands, it often occurs that data for a specific combination of residue and food commodity are scarce, or absent. These problems will certainly multiply if applications in other countries having less measurement data available would be considered. For data-scarce situations a parametric approach to the modelling of residue concentration distributions has been explored. First results indicate a good performance of a binomial-lognormal model, but further research is needed.

1 Introduction

Residues of pesticides or animal drugs and environmental contaminants present a health hazard due to their possible presence in human food. Reliable and accurate quantitative methods are required to assess the dietary exposure to residues and contaminants. Basically, all methods follow the equation

$$Exposure = \frac{\sum consumption \times concentration}{body\ weight}$$

where the summation is over food commodities in the diet which may contain the residue of interest. Consumption is expressed per day, and standardization on body weight then gives the amount of residue per kg body weight per day, which can be easily compared with health norms such as ADI (acceptable daily intake) or ARfD (acute reference dose).

An important distinction is made between *chronic* and *acute* exposure. In the first case variability of consumption and residue levels is less important because of averaging over the long term. In this report we will be concerned with models for acute risk assessment.

In acute risk assessment models the variability in consumption and residue levels cannot be ignored. Therefore a *probabilistic approach* is needed which incorporates the stochastic nature of food consumption and residue concentrations. In this approach, a distribution of food consumption data as well as a distribution of residue data are used.

For both components of the model (consumption data and residue data) a choice can be made between a parametric or a non-parametric approach. In a *parametric approach* the data are modelled with an appropriate distributional form (e.g. lognormal). For modelling the food consumption a multivariate specification is required, due to correlations between consumption of specific products. On the other hand residue concentrations in the various food commodities may be assumed to be independent and therefore can be modelled by univariate distributions. 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.

With random sampling from both the consumption and the residue distribution an exposure value can be calculated. The sampling is repeated many times to generate an *exposure distribution*. The new distribution is analysed to determine statistics for the population, most notably the percentiles in the upper tail.

In this study the probabilistic approach to estimate the dietary intake of residues and contaminants in food is implemented and illustrated using residue data from the KAP database and consumption data from a national survey on food consumption. In chapter 3 a non-parametric approach is used: consumption patterns (with the associated body weight of the sampled individual) and residue values are sampled at random from the available data and merged together to generate a new distribution of exposure values. To assess the risk-exposure, percentiles of the exposure

distribution are estimated. This approach is implemented both in a general-purpose statistical programming language, *Genstat*, and in the currently popular management decision tool *@Risk*, which is an add-in module in *Excel*. A comparison is made between these two implementations.

In chapter 4 a parametric method is described. Consumption patterns are still sampled from the empirical distribution, but 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. 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. 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 was chosen. First, the presence of a non-zero exposure on food products is modelled with a binomial distribution with a parameter p representing the probability of a non-zero residue level. p depends on 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 was selected as being both theoretically sensible and practically useful. In the simplest implementation of this parametric approach residues in each food are modelled separately. However, frequently, data on residues in specific food commodities are sparse or even missing. In those cases, data on similar products may provide the necessary information to base the parameter estimates upon. Pooling of products in product groups to allow joint estimates of parameters is considered in paragraph 4.2. Both versions of the parametric approach have been implemented in *Genstat*. A comparison is made with the non-parametric approach.

2 Available datasets

2.1 Consumption data

In 1992, the second Voedselconsumptiepeiling was carried out among a large number of representative households. On 2 successive days 6218 survey respondents reported their daily consumption of food commodities. These figures were transformed into amounts of raw agricultural products. Respondents were categorized by age and sex among several other characteristics.

2.2 Residue data

The residue data are available from the KAP-database (Oracle), which stores annually more than 200000 records of measurements originating from food monitoring programs for meat, fish, dairy products, vegetables and fruit. In this study attention is restricted to five pesticides which may enter the food chain through vegetables and fruit. Available are data on many products (see Table 1). Missing values in Table 1 may indicate the absence of measurements > 0 or the absence of measurements altogether. In any case, the majority of the results are reported as non-detects ("zeroes") in all cases.

Table 2 gives the summary statistics for Iprodione, Parathion, Chlorothalonil, Pirimicarb and Tolclofos-methyl. For each product the number of non-detects, the number of positive values, the number of products and the average residue concentration of detects and non-detects of all products are reported. The majority of the data is reported as zero. These zeros should be interpreted as non-detectable concentrations. Iprodione is detected the most, the average concentration of all values on 55 products was 0.13 mg/kg

Figure 1 shows the average concentration of all values per product for Iprodione. OAKLEAF LETTUCE (16), LAMB'S LETTUCE (17), TURNIP TOPS/GREENS (18) and OTHER AGRICULTURAL/HORTICULTURAL PRODUCTS (40) have high residue values. Most averages are considerably lower.

Table 1. Total number of measurements per food commodity

residue product	chlr	ipro	para	piri	tolc
“KOUSEBAND” BLACK-EYED PEA	*	7	*	7	*
“RADICCHIO ROSSO”	*	13	*	*	*
“ROODLOF”	*	9	*	*	*
“SPERZIEBOON”	*	101	*	*	102
APPLE	400	400	*	398	*
APRICOT	*	19	*	*	*
AUBERGINE/EGG PLANT	67	*	*	68	*
BEAN (BROWN/YELLOW/WHITE)	*	*	*	8	*
BEAN, (SCARLET/STRING/FRENCH)	*	161	*	161	*
BLACK RADISH	*	*	29	*	29
BLACKBERRY	*	58	*	55	*
BLEACH-CELERY	66	66	66	66	66
BLUE BERRY	*	18	*	15	*
BROAD BEAN	8	*	*	8	*
BROCCOLI	*	*	*	62	*
BRUSSELS SPROUTS	*	41	*	*	*
CABBAGE LETTUCE, COS LETTUCE	670	670	668	668	668
CANTHARELLE	*	*	*	1	*
CARROT	*	125	*	124	124
CAULIFLOWER	*	126	*	127	*
CELERIAC	39	39	39	*	*
CELERY	120	120	121	121	121
CHERVIL	*	8	*	*	*
CHICORY	*	105	*	103	*
CHINESE CABBAGE	112	112	116	116	116
CUCUMBER	245	245	*	249	*
CURLY LETTUCE	*	16	*	16	16
CURRANT (RED, WHITE, BLACK)	*	142	126	126	*
ENDIVE	367	367	366	366	366
FENNEL (FRESH)	*	50	49	*	49
GHERKIN/PICKLE	*	10	*	*	*
GRAPE	198	198	196	*	*
GREEN PEA (FRESH)	22	*	*	*	*
ICEBERG LETTUCE	176	176	178	178	178
KIWI FRUIT	*	61	*	*	*
LAMB’S LETTUCE	*	53	53	*	53
LEAF LETTUCE	*	2	*	2	2
LEEK	68	*	102	*	102
LEGUME	33	*	*	33	*
LOLLO ROSSA	89	89	89	89	89
MANDARIJN, CLEMENTINE	*	*	112	*	*
MELON	95	*	*	*	*
MIXED VEGETABLES	38	38	*	43	43
MUSHROOM	98	*	*	*	*
NECTARINE	*	65	*	*	*
OAKLEAF LETTUCE	*	53	53	53	53
ONION (SMALL)	24	24	*	*	*
ORANGE	*	*	335	335	*
OTHER AGR./HORTICULT. PRODUCTS	*	54	*	54	*
OTHER FRUIT, NUTS	*	18	*	*	*
OXHEART/CONICAL CABBAGE	*	53	*	53	*
PAC-CHOY	*	40	*	41	41
PARSLEY, ROOTED PARSLEY	84	*	85	85	85
PASSION FRUIT	17	*	*	*	*
PEACH	*	39	*	*	*
PEAR	*	90	*	92	*
PEPPER	117	117	119	*	*
PLUM, INCLUDING DAMSON	*	81	*	85	*
POTATOES	*	149	*	*	*
PUMPKIN, COURGETTE	82	82	*	88	*
PURSLANE	*	30	*	30	30
RADISH	*	167	167	*	167
RASPBERRY	86	86	68	68	68
SPINACH	172	172	170	170	170
STRAWBERRY	978	978	779	779	779
SWEET CHERRY	*	44	*	*	*
SWEET PEPPER	488	488	*	489	*
TOMATO	372	372	*	377	377
TURNIP TOPS/GREENS	*	28	*	*	28
WINTER CARROT	*	40	*	*	*

Table 2. Summary statistics: number of observations, number of products, average residue concentration of all products, percentage of detects (non-zero's) and the average residue concentration of the non-zero's

residue	number of observations	number of products	average concentration (mg/kg)	percentage of detects (%)	average concentration in detects only (mg/kg)
Iprodione	6915	55	0.13	14	0.83
Parathion	4086	23	0.01	2	0.27
Chlorothalonil	5307	29	0.05	2	0.55
Pirimicarb	5831	41	0.02	6	0.24
Tolclofos	3922	26	0.02	12	0.13

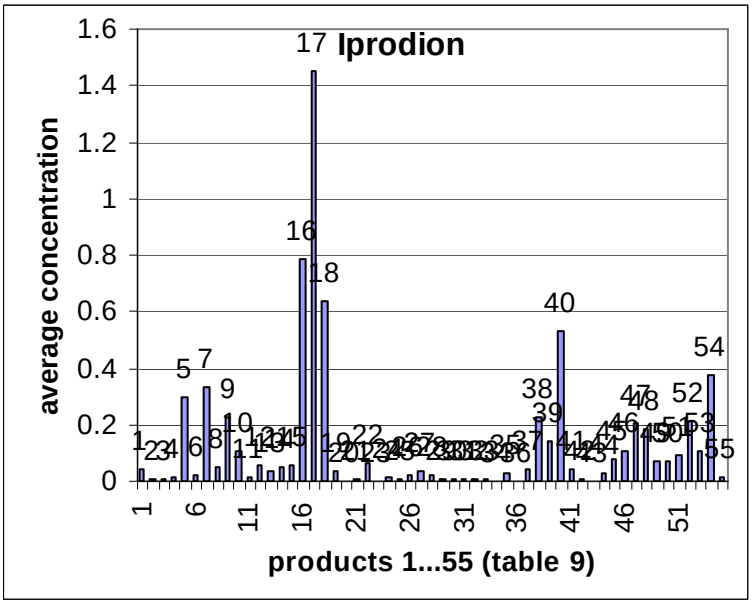

Figure 1. Average concentration (mg/kg) of detects and non-detects of Iprodione for all products

3 Non-parametric approach

Use is made of *Oracle* databases containing the food consumption and residue concentration data. For each residue under investigation (e.g. Iprodione in the example below), selections are made from these databases to provide six ASCII files:

Table 3. Files and records

Personen.lis	Individual number, age (years), weight (kg) (of respondents in Food Consumption Survey, this file is the same for all residues)
Ipro_con.lis	Individual number, day (1 or 2), product number, amount consumed (g) (this file may exclude products that never contain the residue)
Ipro_geh.lis	Product number, measured concentration (mg/kg) (File has a line for each non-zero measurement)
Ipro_nge.lis	Product number, total number of measurements per product (including zeroes)
Ipro_prd.lis	Product number, product name
Ipro_sto.lis	Residue number, residue name (this file contains only one line)

Product and residue number have a hierarchical structure, allowing to consider product groups or residue groups by ignoring some of the digits. See Appendix for more information.

The ASCII files are used in both the Genstat and @Risk implementations described below in detail. In both cases individual daily consumption patterns are randomly selected from the *con* file, and residue concentrations are sampled from the *geh* and *nge* files. The non-parametric exposure distribution is obtained by repeatedly multiplying the sampled values together, summing over the food products, and dividing by body weight which is retrieved from the *personen* file. Product and residue names are retrieved from the *prd* and *sto* files to obtain a more readable output.

3.1 Genstat implementation

The Genstat program is fast, because sampling is done in parallel data structures. Let n be the chosen number of simulations, and k the number of food products. Then the program selects a simple random sample of n individual numbers and a simple random sample of n day numbers (1 or 2). The selection of individuals may be restricted to a specified range of ages (for example, only children from 0 to 4 years). A typical value of n may be 100,000. Note that each of the 6218 individuals is likely to occur many times in the sample. For each iuxtaposed combination of sampled individual and sampled day the consumption data are retrieved from the *con* file, and stored in an $n \times k$ matrix.

Another $n \times k$ matrix is constructed to contain simulated concentration data. For all k products the total number of measurements (t) and the number of non-zero measurements (w) is determined from the *nge* and *geh* files. Then random index numbers (i) between 1 and t are sampled for each cell of the matrix. If $i \leq w$, then the i^{th} value for this product is selected from the *geh* file. If $i > w$, then a value 0 is inserted in the concentration matrix.

Both $n \times k$ matrices are now multiplied elementwise, and all values are divided by 1000, because consumption is in g, but exposure in mg. Summing over the k products and dividing by the n body weights corresponding with the n selected individuals then gives the simulated exposure distribution as a vector of n values. Relevant percentiles can be obtained from this vector.

3.2 @Risk implementation

@Risk is a simulation add-in for Excel and adds Monte Carlo simulations to spreadsheets. Uncertain values in the spreadsheet are replaced by @Risk or user-defined probability distribution functions. Spreadsheets are recalculated sequentially 10.000 – 50.000 of times, each time sampling random values from the @Risk functions. The sequential nature of the spreadsheet recalculations makes the @Risk implementation much slower than the Genstat program (hours instead of minutes), thereby limiting the practical number of simulations.

The result is a distribution of possible outcomes, which again can be investigated for the relevant percentiles. The Monte Carlo simulation in @Risk can carried out either by simple random sampling or by Latin Hypercube sampling. The latter method is in theory more efficient, and was therefore used in this study.

The practical implementation of risk analysis in Excel and @Risk is an Excel worksheet *risico.* This worksheet contains references to 6 worksheets in the same Excel workbook, which contain copies of the 6 above-mentioned ASCII files. The worksheet *risico* makes use calculations involving the @Risk functions *RiskDuniform* and *RiskDiscrete*, which are recognized by @Risk during the simulation, and used for sampling the data in the other sheets. At any time the worksheet *risico* shows the results of one simulation (see Table 4).

Table 4. Example of @Risk worksheet *risico* (part) showing one simulation for Iprodione. A slight exposure of a 20-year old individual is in this case mainly due to consumption of CARROT (WORTEL) (and in a lesser extent CURRANT (AALBES)). This person is lucky not to eat highly contaminated BLACKBERRY (BRAAM), but instead lots of clean APPLE (APPEL) and GRAPE (DRUIF).

RISICOANALYSE IPRODION (=GLYCOFEEN)					
Dag	Persoon	Leeftijd	Gewicht		BELASTING
1	358954	20	84	Totaal:	0.013486
		jaar	kg		microg/kg
		CONSUMPTIE (g)	GEHALTE (mg/kg)		BELASTING (microg/kg)
BOON, (PRONK/SLA/SNIJBOON)		0	0		0
SPERZIEBOON		0	0		0
WITLOF		0	0		0
ROODLOF		0	0		0
ANDIJVIE		0	0		0
IJSBERGSLA		0	0		0
KROPSLA, BINDSLA		0	0		0
KRULSLA		0	0		0
LOLLO ROSSA		0	0		0
PLUKSLA		0	0.21		0
SELDERIJ		0.19	0		0
SPINAZIE		0	0		0
KERVEL		0	0		0
POSTELEIN		0	0		0
RADICCHIO ROSSO		0	0		0
EIKEBLADSLA		0	0		0
VELDSLA		0	0		0
RAAPSTELEN, RUCOLA		0	0		0
BLEEKSELDERIJ		0	0		0
BLOEMKOOL		10.097	0		0
SPRUITKOOL		0	0		0
CHINESE KOOL		0	0		0
SPITSKOOL		0	0		0
BOSUI		0	0		0
VENKEL (VERS)		0	0		0
AARDAPPELEN		13.856	0		0
WINTERWORTEL		0	0		0
WORTEL		11.48	0.09		0.0123
RADIJS		0	0		0
KNOLSELDERIJ		0	0		0
KOMKOMMER		0	0		0
TOMAAT		13.802	0		0
PAPRIKA		0.96	0		0
POMPOEN, COURGETTE		0	0		0
PEPERS		0	0		0
AUGURK		0	0		0
KOUSEBAND		0	0		0
PAKSOI		0	0		0
GEMENGDE GROENTEN		0	0		0
OV. LAND- EN TUINBOUWPRODUKTEN		0	0		0
KIWI		0	0		0
APPEL		270.48	0		0
PEER		1.664	0		0
ABRIKOOS		1.636	0		0
NECTARINE		0	0		0
PERZIK		1.64	0		0
PRUIM, INCL KWETS		0	0.02		0
KERS		1.648	0		0
DRUIF		252.35	0		0
AARDBEI		1.652	0		0
FRAMBOOS		1.66	0		0
BRAAM		0	1.9		0
BLAUWE BES		0	0		0
AALBES (ROOD, WIT, ZWART)		1.66	0.06		0.001186
OV. FRUIT, NOTEN		0	0		0

3.3 Results

We first give some detailed results for Iprodione, and then a summary of the calculated exposure percentiles for all 5 residues.

Table 5 summarizes the simulation results relating to Iprodione using Genstat and @Risk. The main results are the percentiles of the intake distribution. For example, a 99 % percentile of 4 μg/kg/day means that an intake of at least this level is expected for 1 out of every 100 intakes, or 10 out of every 1000 intakes, or 100 out of every 10000 intakes , etc.

Iprodione is measured on 55 products. Only 937 residues were positive (14%), the number of non-detects was 5978 (86%). In total 6218 persons were surveyed on 2 successive days, giving an incidence matrix with 87117 values. The number of iterations was 50.000. A simulation run with @Risk took 2h.9', Genstat completed the task within a 2 minutes. Between Genstat and @Risk only minor differences occur, no more than between repeated simulations with any one of the programs. Results are relatively stable for the estimates of the 95^{th} , 98^{th}, 99^{th} percentile of exposure to Iprodione. Discrepancies occur at the higher percentiles, e.g. estimates of the 99.99^{th} percentile range from 44 to 74 μg/kg/day.

Table 5. Estimates of percentiles of exposure to Iprodione (μg/kg/day) for simulations with Genstat and @Risk using non-parametric approach (50000 iterations)

Iprodione	Percentile	95 %	98 %	99 %	99.5 %	99.9 %	99.99 %
Genstat		0.66	2.1	4.3	7.0	18	64
		0.62	2.2	4.5	8.4	22	74
		0.64	2.0	4.0	6.6	16	44
@Risk		0.64	2.1	4.2	7.0	17	47

Figure 2 shows the empirical frequency distribution for exposure to Iprodione (50000 iterations). Table 6 shows the products which are responsible for the 10 highest simulated intakes, with ENDIVE as a main source of very high intakes.

Table 7 shows estimated percentiles for all 5 example substances.

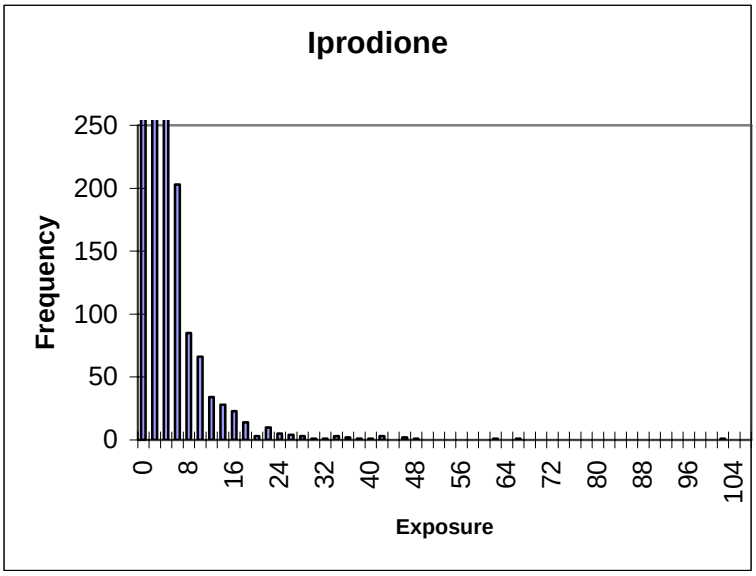

Figure 2. Upper tail of the empirical frequency distribution for 50000 Iprodione exposure simulations. The first three bars (cut off at 250) represent zero exposures (25541 values) and positive values ≤ 2 μg/kg/day (23469 values) and values >2 and ≤ 4 μg/kg/day (494 values).

Table 6. Non-parametric approach. Top 10 (0.02 %) of 50000 simulated total Iprodione intakes (μg/kg/day), traced to responsible products (excluding contributions < 0.5 μg/kg/day).

Total intake	100	66	61	46	45	44	42	41	41	40
ENDIVE	100		61	46	45	44				
CABBAGE LETTUCE		66							1	
SPINACH							42			
LAMB'S LETTUCE								41	40	40

Table 7. Percentiles of exposure distributions for five residues in μg/kg/day. Non-parametric approach, Genstat implementation, 50000 iterations.

Percentile	95 %	98 %	99 %	99.5 %	99.9 %	99.99 %
Residue						
Chlorothalonil	0.02	0.10	0.20	0.39	1.6	7.5
Iprodione	0.64	2.0	4.0	6.6	16	44
Parathion	0.001	0.02	0.07	0.18	1.9	26
Pirimicarb	0.03	0.14	0.41	0.88	3.1	9.4
Tolclofos-methyl	0	0.008	0.08	0.24	1.1	4.2

4 Parametric approach

The residue data on commodities contain many non-detects, which appear as zeros in the database. Therefore, it was decided to proceed the parametric approach in two steps: a binomial distribution was fitted to the zero and non-zero data, giving probability *p,* while a lognormal model was used to generate the residue distribution of the non-zero data. Parameter *p* of the binomial was estimated as the fraction of detects, the parameters, μ and $\bullet$ of the lognormal were based on the logtransformed non-zero residues.

4.1 Comparing distributions

BestFit is a decision tool, which can be linked to Excel to fit data to more than 30 distribution types. It performs statistical tests to compare quality of fit and ranks distributions by three goodness-of-fit statistics. In this study we used the Anderson-Darling test which is similar to the Kolmogorov-Smirnov test, but places more emphasis on the tail values. All tests are very sensitive to the number of values. Graphs are used to assess visually how well distributions agree with the input data. Both test statistics as graphs should be used in interpreting the results.

BestFit was used to explore the non-zero residue values to find which of the distribution types fits best. Products with at least 30 positive measurement values were taken to explore which distributional type was suitable. Table 8 summarizes accepted test-results according to the Anderson-Darling goodness-of-fit statistic. Distributions that are not in the table are rejected by all goodness-of-fit statistics. The lognormal and the Pearson VI turned out to have an adequate fit for Iprodione content in all five products.

Table 8. Test results for fitting distributions to positive Iprodione contents in ENDIVE, CABBAGE LETTUCE, CURRANT, CARROT and STRAWBERRY. n = sample size, acc = acceptable distribution according to Anderson-Darling test (95 % significance test). In parentheses are the ranks of the best fitting distributions. Calculations with BestFit.

	ENDIVE	CABBAGE LETTUCE	CURRANT	CARROT	STRAWBERRY
distributions	n=92	n=286	n=30	n=36	n=169
Lognormal	acc (1)	acc (3)	acc (4)	acc (4)	acc (1)
Pearson V	acc (2)	acc (2)	acc (6)	acc (9)	
Pearson VI	acc (4)		acc (2)	acc (7)	acc (2)
Weibull			acc (1)	acc (6)	
InverseGaus	acc (3)	acc (1)		acc (3)	acc (3)
Chisq	acc (5)				
Beta			acc (5)	acc (8)	
Gamma			acc (3)	acc (1)	
Rayleigh				acc (2)	
Triang				acc (10)	
ExtremeValue				acc (5)	
Logistic					
Pareto			acc (7)		
Normal					

BestFit ranks distributions by the value of the test statistic. The ranking is represented by the number in parentheses. For all products lognormal has a high ranking. Figure 3 displays the corresponding input and fitted lognormal distributions. For comparison also the fitted normal distribution is added to each figure.

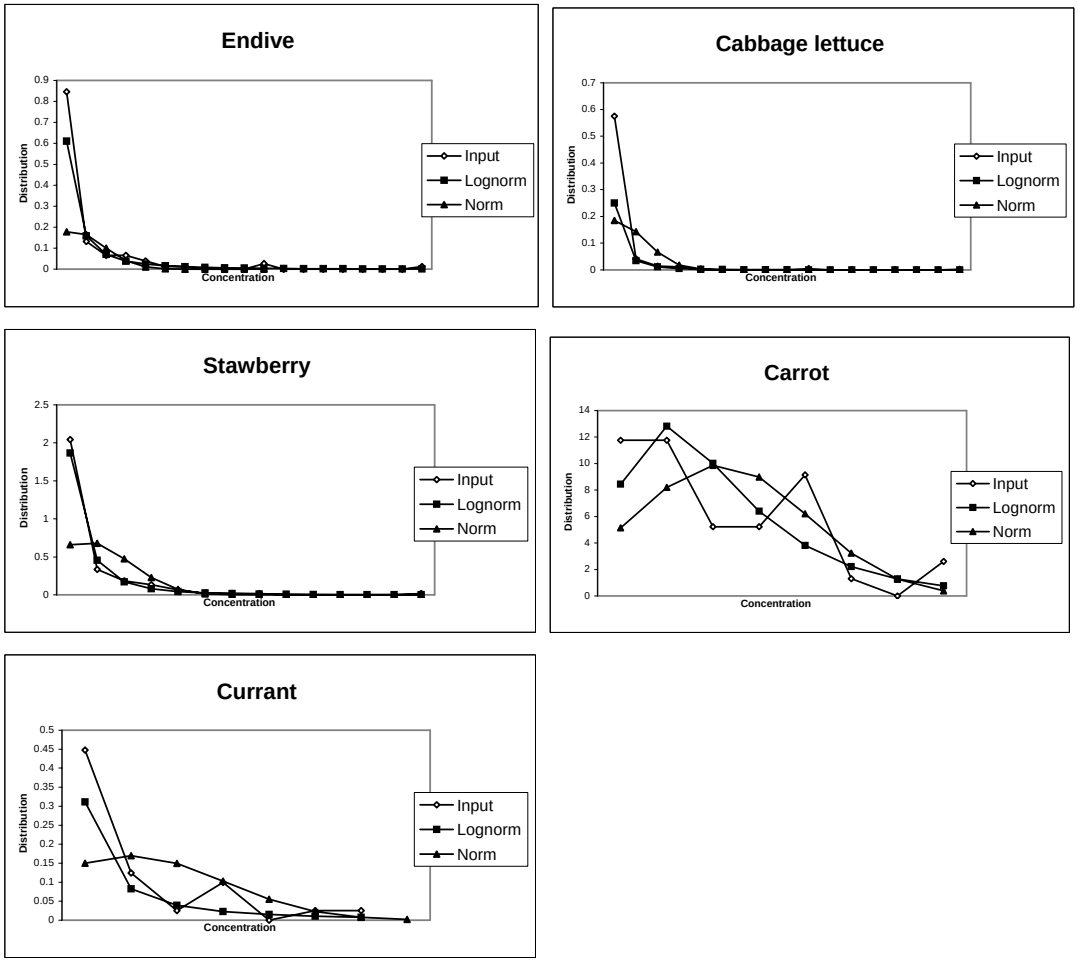

Figure 3. Distribution of non-zero Iprodione concentrations for ENDIVE, CABBAGE LETTUCE, STRAWBERRY, CARROT and CURRANT. Input = empirical frequency distribution. Lognormal and Normal distributions were fitted to the Input data. These graphs illustrate the results of Table 8**: satisfactory fit of the lognormal and unsatisfactory fit of the normal distribution.**

Since residue data are positive, positively skewed and originate by mechanisms which generate the lognormal distribution under a variety of biological circumstances (Crow & Shimizu, 1988) the lognormal was chosen to proceed with. Ease of interpretation comprised another significant reason to choose the lognormal. Therefore we propose the lognormal distribution as a general model for positive concentrations, at least in those situations where it is not contradicted by the data. Of course, in future research with more data available, the choice of distributional form should be reinvestigated.

4.2 Sparse or missing data: pooled estimates for the variance and mean

Estimation of the sample variance and/or mean are often hampered by sparse or even missing data. In those cases, rearrangement of products into groups may give sufficient data to base estimates upon. 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 and is applied to the Iprodione data as an example:

1. Estimate the variance and sample mean for each product, giving sigma1, mu1 and df1, see Table 9. Notice, that in some cases the variance is missing because only one measurement is available (e.g. “ROODLOF”, 10801, sigma1 = *). Then, products are assigned to productgroups. Within each productgroup products are marked to indicate whether the use of residues is allowed or not. The homogeneity of variances in different (marked) productgroups can be assessed using Bartlett's test (Snedecor & Cochran, 1980). The test statistic determines whether variances are to be pooled automatically (p>0.05) or not (p• 0.05). In the latter case, products are assigned to subgroups (within productgroups) by hand and the homogeneity of variances is tested again. For homogeneous groups, variances are pooled within productgroups. This process of assigning products to subgroups is repeated until all groups have homogeneous variances. After pooling the variances, an overall test for differences of means is performed, based on analysis of variance. Means are pooled automatically if the probability is > 0.05. If not, the original means are maintained. Table 9 shows the above procedure. The variances of productgroup 10701* (BEAN AND “SPERZIEBOON”) are pooled automatically: sigma2 = 1.31, df2 = 12 (= 5 + 7). The probability of the test for differences of means is *p* >0.05, so means are pooled automatically as well: mu2 = -1.66. The variances of productgroup 10801* (CHICORY, ENDIVE, …, LOLLO ROSSA, RADICCHIO ROSSO, …, LAMB’S LETTUCE) are pooled auto-matically: sigma2 = 1.48, df2 = 439 (= 3 + 91 + … + 16), but now, means significantly differ. The variances for productgroup 10904* (STRAWBERRY, …, CURRANT) are heterogeneous, so this group is rearranged by hand into two new subgroups: 1) STRAWBERRY (1.14) and BLACKBERRY (1.15), and 2) RASPBERRY (1.73), BLUE BERRY (1.83) and CURRANT (1.87). Now, variances within subgroups are homogeneous and are pooled, yielding sigma2 = 1.14 for the first and 1.84 for the second subgroup. The means for the second group are pooled automatically: mu2 = -0.76, the means for STRAWBERRY and BLACKBERRY are maintained: mu2 = –1.57 and –0.89, respectively. Missing variances, e.g. “ROODLOF”, are replaced by the pooled variance of the productgroup (10801): sigma2 = 1.28. The missing variance of “KOUSEBAND” (10889) remains missing: sigma2 = *, because no (pooled) variance is available in this productgroup. Optional is step 4.
2. 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) 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. If the variance is replaced for (sub)groups with two or more members, a test for differences of means is performed. Means are pooled automatically if p>0.05, if not, the original means are maintained. Table 9 shows how the above is implemented. E.g. BRUSSELS SPROUTS and OXHEART/CONICAL CABBAGE (10802) have less than 10 df with sigma2 = 1.14. The variances are tested against the corrected overall variance, the probability is > 0.05, so their variances are replaced: sigma3 = 1.36 and df3 = 882. The means, -2.70 and –2.30 are tested with p > 0.05 and are pooled automatically: mu3 = -2.57. Conversely, the variances of ONION (SMALL) and

FENNEL (10803*) are not replaced so sigma3 = 0.14. The missing variance of "KOUSEBAND" (10889) is replaced: sigma3 = 1.36 with df3 = 882.

3. After carrying out the above pooling process, there still remain products with less than 10 df. These products are considered again. The variances are judged visually and assigned by hand to one or more of the products with approximately the same value for the (pooled) variance, After testing the variances, the variances are pooled again, replacing the variance based on < 10 df with the pooled one. Testing for differences of means is performed and for those cases where p>0.05, means are also pooled. E.g. PAC-CHOY (10889*) has less than 7 df and is assigned to 10903. The pooled variance: sigma3 = 1.28 with df3 is 37 (= 30+7), the original mean is maintained. ONION (SMALL) and VENKEL are not assigned to any group, so the original variance is kept: sigma3 = 0.14.
4. Step 4 is optionally for those cases where variances are pooled, but means are not. Products may be rearranged into (sub)productgroups based on similarity of their means. Then, pooled means are calculated replacing the original ones. E.g. productgroup 10801* (CHICORY, ENDIVE, …, LOLLO ROSSA, RADICCHIO ROSSO, OAKLEAF LETTUCE, LAMB'S LETTUCE) has a pooled variance: sigma3 = 1.48 but the means are original. Visually, with in parentheses the estimate of the mean, CHICORY (-2.69), ICEBERG LETTUCE (-1.92), CABBAGE LETTUCE (-1.44) and CURLY LETTUCE (-2.14) are assigned to a subgroup, ENDIVE (-0.91), LOLLO ROSSA (-0.99), RADICCHIO ROSSO (-0.36) and OAKLEAF LETTUCE (-0.06) to a second group while LAMB'S LETTUCE (0.80) forms a single group. After pooling, the new means, mu4, for the three subgroups are: –1.48, -0.83 and 0.80, respectively.

Table 9. Standard deviation (sigma), mean (mu) and degrees of freedom (df) in different pooling steps. The asterisk indicates that the use of Iprodione on the product is allowed.

Product	Product-group	sigma1	mu1	df1	sigma2	mu2	df2	sigma3	mu3	mu4	df3
1 BEAN	10701*	1.60	-1.17	7	1.31	-1.66	12	1.31	-1.66	-1.66	12
2 "SPERZIEBOON"	10701*	0.75	-2.33	5	1.31	-1.66	12	1.31	-1.66	-1.66	12
3 CHICORY	10801*	1.38	-2.69	3	1.48	-2.69	439	1.48	-2.69	-1.48	439
4 "ROODLOF"	10801	*	-2.30	0	1.28	-0.58	16	1.28	-0.58	-0.58	16
5 ENDIVE	10801*	1.52	-0.91	91	1.48	-0.91	439	1.48	-0.91	-0.83	439
6 ICEBERG LETTUC	10801*	1.65	-1.92	7	1.48	-1.92	439	1.48	-1.92	-1.48	439
7 CABBAGE LETTUCE	10801*	1.46	-1.44	285	1.48	-1.44	439	1.48	-1.44	-1.48	439
8 CURLY LETTUCE	10801*	1.08	-2.14	3	1.48	-2.14	439	1.48	-2.14	-1.48	439
9 LOLLO ROSSA	10801*	1.53	-0.99	21	1.48	-0.99	439	1.48	-0.99	-0.83	439
10 LEAF LETTUCE	10801	*	-1.56	0	1.28	-0.58	16	1.28	-0.58	-0.58	16
11 CELERY	10801	1.76	-1.27	2	1.28	-0.58	16	1.28	-0.58	-0.58	16
12 SPINACH	10801	1.18	-0.57	9	1.28	-0.58	16	1.28	-0.58	-0.58	16
13 CHERVIL	10801	*	-1.24	0	1.28	-0.58	16	1.28	-0.58	-0.58	16
14 PURSLANE	10801	*	0.41	0	1.28	-0.58	16	1.28	-0.58	-0.58	16
15 RADICCHIO ROSSO	10801*	*	-0.36	0	1.48	-0.36	439	1.48	-0.36	-0.83	439
16 OAKLEAF LETTUCE	10801*	1.65	-0.06	13	1.48	-0.06	439	1.48	-0.06	-0.83	439
17 LAMB'S LETTUCE	10801*	1.25	0.80	16	1.48	0.80	439	1.48	0.80	0.80	439
18 TURNIP TOPS/GREE	10801	1.19	1.30	2	1.28	-0.58	16	1.28	-0.58	-0.58	16

19 BLEACH-CELERY	10801	1.23	-0.88	3	1.28	-0.58	16	1.28	-0.58	-0.58	16
20 CAULIFLOWER	10802*	*	-1.83	0	1.62	-1.83	20	1.62	-1.83	-1.83	20
21 BRUSSELS SPROUT	10802	1.14	-2.70	1	1.14	-2.70	1	1.36	-2.57	-2.57	882
22 CHINESE CABBAGE	10802*	1.62	-2.32	20	1.62	-2.32	20	1.62	-2.32	-2.32	20
23 OXHEART/CONICAL	10802	*	-2.30	0	1.14	-2.30	1	1.36	-2.57	-2.57	882
24 ONION (SMALL)	10803*	0.07	-1.66	1	0.14	-1.66	3	0.14	-1.66	-2.09	3
25 FENNEL	10803*	0.16	-2.38	2	0.14	-2.38	3	0.14	-2.38	-2.09	3
26 POTATO	10804	0.62	0.19	1	0.59	0.19	50	0.59	0.19	0.19	50
27 WINTER CARROT	10804	0.62	-2.55	13	0.59	-2.55	50	0.59	-2.55	-2.64	50
28 CARROT	10804	0.54	-2.71	35	0.59	-2.71	50	0.59	-2.71	-2.64	50
29 RADISH	10804*	1.52	-2.91	5	1.52	-2.91	5	1.36	-2.91	-2.91	882
30 CELERIAC	10804	1.31	-2.07	1	0.59	-2.07	50	0.59	-2.07	-2.64	50
31 CUCUMBER	10805*	0.80	-1.55	7	0.99	-2.22	26	0.99	-2.22	-2.22	26
32 TOMATO	10805*	0.88	-2.50	13	0.99	-2.22	26	0.99	-2.22	-2.22	26
33 SWEET PEPPER	10805*	1.33	-2.19	6	0.99	-2.22	26	0.99	-2.22	-2.22	26
34 PUMPKIN,	10805*	*	-2.53	0	0.99	-2.22	26	0.99	-2.22	-2.22	26
35 PEPPER	10805	1.23	-0.94	4	1.23	-0.94	4	1.36	-0.94	-0.94	882
36 GHERKIN/PICKLE	10805*	*	-3.51	0	0.99	-2.22	26	0.99	-2.22	-2.22	26
37 “KOUSEBAND”	10889	*	-1.27	0	*	-1.27	0	1.36	-1.27	-1.27	882
38 PAC-CHOY	10889*	1.29	-0.48	7	1.29	-0.48	7	1.28	-0.48	-0.48	37
39 MIXED VEGETABLE	10890	1.37	0.09	2	2.22	-0.45	6	1.36	-0.45	-0.45	882
40 OTHER AGR./HORTI	10890	2.54	-0.77	4	2.22	-0.45	6	1.36	-0.45	-0.45	882
41 KIWI FRUIT	10901	1.96	-0.96	2	1.96	-0.96	2	1.36	-0.96	-0.96	882
42 APPLE	10902	2.02	-1.59	3	2.02	-1.59	3	1.36	-1.97	-1.97	882
43 PEAR	10902	*	-3.51	0	2.02	-3.51	3	1.36	-1.97	-1.97	882
44 APRICOT	10903	1.55	-1.71	1	1.26	-0.83	30	1.26	-0.83	-0.83	30
45 NECTARIN	10903	1.08	-0.97	8	1.26	-0.83	30	1.26	-0.83	-0.83	30
46 PEACH	10903	1.18	-0.74	5	1.26	-0.83	30	1.26	-0.83	-0.83	30
47 PLUM, INCLUDING	10903	1.34	-0.69	16	1.26	-0.83	30	1.26	-0.83	-0.83	30
48 SWEET CHERRY	10903*	0.89	-0.69	11	0.89	-0.69	11	0.89	-0.69	-0.69	11
49 GRAPE	10904	1.14	-1.06	24	1.14	-1.06	24	1.14	-1.06	-1.06	24
50 STRAWBERRY	10904*	1.14	-1.57	168	1.14	-1.57	184	1.14	-1.57	-1.51	184
51 RASPBERRY	10904*	1.73	-1.04	8	1.84	-0.76	39	1.84	-0.76	-0.76	39
52 BLACKBERRY	10904*	1.15	-0.89	16	1.14	-0.89	184	1.14	-0.89	-1.51	184
53 BLUE BERRY	10904*	1.83	-1.24	2	1.84	-0.76	39	1.84	-0.76	-0.76	39
54 CURRANT	10904*	1.87	-0.62	29	1.84	-0.76	39	1.84	-0.76	-0.76	39
55 OTHER FRUIT, NUT	10990	*	-1.51	0	*	-1.51	0	1.36	-1.51	-1.51	882

4.3 Results of parametric approach

The simulations were performed using Genstat. Table 10 summarizes the results of the Monte Carlo study using parametric distributions. The number of iterations was 50,000 and a simulation run took approximately 13 minutes. The estimates are based on the parameter estimates of Table 9, which is produced applying step 1, 2 and 3 and the optional step 4.

Table 10. Estimates of percentiles of exposure to Iprodione using parametric and non-parametric approach. Simulations with Genstat (50000 iterations).

Iprodione	Percentile	95 %	98 %	99 %	99.5 %	99.9 %	99.99 %
Parametric	Genstat	0.68	2.4	4.6	7.7	23	66
		0.71	2.3	4.6	7.7	20	55
		0.71	2.1	4.1	7.1	20	58
Nonparametric (results copied from Table 5)	Genstat	0.66	2.1	4.3	7.0	18	64
		0.62	2.2	4.5	8.4	22	74
		0.64	2.0	4.0	6.6	16	44

For comparison the results of the nonparametric approach have also been included in Table 10. It can be seen that the percentiles yield estimates of risk exposure that are in the same range, although, vaguely the parametric equivalent seems to give slightly less variable results, and provide somewhat higher estimates of the 95th percentile. However, such differences should be investigated more fully in future research.

The interpretation of the 99.99 % percentile is that only 5 intakes out of 50000 are expected to be higher than about 60 μg/kg/day. Table 11 shows which products are responsible for the 10 highest simulated intakes in one of the parametric simulations. Just as in the non-parametric approach, ENDIVE turns out to be the main source of very high intakes.

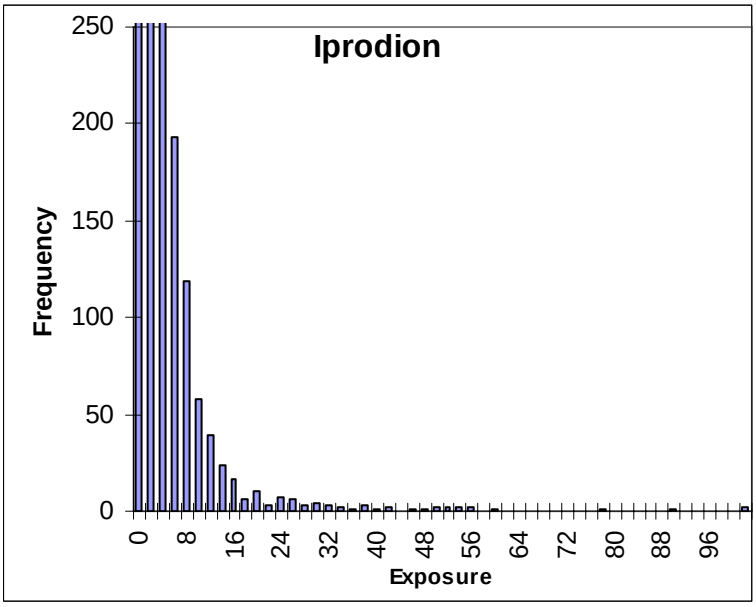

Figure 4. Upper tail of the parametric frequency distribution for Iprodione exposure. The first three bars (cut off at 250) represent zero exposures (25659 values) and positive values ≤ 2 μg/kg/day (23285 values) and values >2 and ≤ 4 . μg/kg/day (540 values).

Table 11. Parametric approach. Top 10 (0.02 %) of 50000 simulated total Iprodione intakes (μg/kg/day), traced to responsible products (excluding contributions < 0.5 μg/kg/day).

Total intake	280	110	88	78	60	56	54	54	52	52
"SPERZIEBOON"					56					
ENDIVE	280	110		78				54	52	
SPINACH							54			
LAMB'S LETTUCE										52
POTATO					4					
STRAWBERRY						2				
BLUE BERRY			88							
APPLE						54				

5 Discussion

How many data are required for a sensible calculation of upper-tail percentiles in the exposure distribution? This report has described a non-parametric probabilistic approach which needs sufficient representative measurements of the residue under consideration for *each* food commodity which might contribute to residue intake in substantial amounts. In the absence of a parametric model 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 95^{th} percentile (each measurement represents 5 % of the distribution), and at least 100 measurements to estimate the 99^{th} percentile (each measurement represents 1 %). More generally, the number of measurements per food commodity (n) should at least equal 1/(1-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.

A cautionary note is that enlarging the number of simulated intakes (from 50,000 to 500,000 say) may stabilize the results obtained from the data set, but does never compensate for incidental extreme values or lack of detects due to small size of the data. Said otherwise: random errors in the data (likely in small data sets) behave as systematic effects in repeated analyses of the same data set, and thus may give a false impression of reliability. In future research the variability of the risk percentile estimates as a function of both sample size and simulation size should be investigated more fully, e.g. using bootstrap methodology.

In the practical examples of this study there were generally enough observations per product to allow a non-parametric approach for the 95^{th} percentile. Only for some less consumed products (APRICOT, GHERKIN/PICKLE, BLUE BERRY, CANTHARELLE, CHERVIL, "KOUSEBAND"/BLACK-EYED PEA, CURLY LETTUCE, OTHER FRUIT & NUTS, PASSION FRUIT, LEAF LETTUCE, "RADICCHIO ROSSO", "ROODLOF", BEAN, BROAD BEAN) the number of measurements was below 20. Assuming that no important food components are missing from Table 1 and that in the table missing values for important food products represent at least 20 non-detect measurements, the data seem suitable for the risk analysis.

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). Two options can be chosen, or combined:

1. *Parametric* modelling: 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 measurements > 0. Figure 5 shows which approach could be best used depending on the total number of measurements and the number of positive measurements. In principle, such a choice could be made separately for each food commodity.
2. *Grouping of products*: this enlarges the number of measurements per group. We must assume now that the residue distributions are the same for the grouped products. In the case of parametric modelling the assumptions of equality can be

restricted to a subset of the parameters (in the chosen binomial-lognormal model: detect probability, lognormal mean, lognormal standard deviation).

In this study only one simple parametric approach has been implemented, using a binomial-lognormal model for all products, and pooling the lognormal parameters where possible. Further studies are needed to investigate a flexible procedure for dietary risk analysis which would adapt to the varying practical database conditions

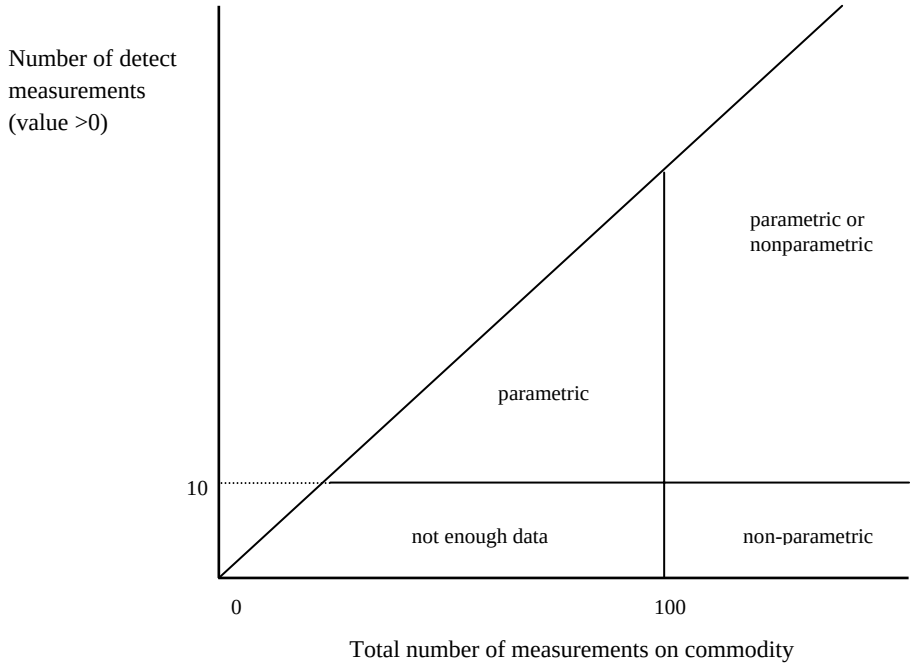

Figure 5. Use of non-parametric or parametric modelling for estimating 99 % exposure percentile in relation to sample size and number of positive measurements.

6 Conclusions

The results of the project described in this report are:

- A Genstat program and an Excel worksheet (to be used with @Risk) both implementing the nonparametric method of dietary acute risk assessment. The Genstat implementation allows for a selection on age of individuals in the database. With the @Risk implementation Latin hypercube sampling can be used instead of random sampling. See Chapter 3.
- A Genstat program implementing a parametric method of dietary acute risk assessment. The parametric method employs a binomial/lognormal model for the residue concentrations. There are possibilities for automatic or user-defined pooling of parameters over products. See Chapter 4.
- A comparison of distributions to be used in the parametric approach using the program Bestfit. See section 4.1.
- Application of the nonparametric method to datasets of five example residues (chlorotalonil, Iprodione, Parathion, Pirimicarb and Tolclofos-methyl) and application of the parametric method to the data of one residue (Iprodione). In all cases upper-tail percentiles were calculated, and for Iprodione details were given on the contributions of specific products to the upper tail. See sections 3.3 and 4.3.
- A discussion of the potentials of both the parametric and nonparametric approach for use in situations where data are not abundant. See Chapter 5.

7 Literature

@Risk (1996). Advanced risk analysis for spreadsheets, Windows version. Pallisade Corporation, Newfield, NY, USA.

Bestfit.(1997). Probability distribution fitting for Windows. Pallisade Corporation, Newfield, NY, USA.

Crow, E.L. and K. Shimizu (1988). Lognormal Distributions: theory and applications, p8. Marcel Dekker, INC, New York.

Genstat 5 Committee (1993). Genstat 5 Release 4.1 (Fourth Edition) Reference Summary. Clarendon Press, Oxford.

Petersen, B.J., L.M. Barraj, L.R. Muenz and S.L. Harrison (1994). An alternative approach to dietary exposure assessment. Risk Analysis, Vol. 14, No. 6.

Snedecor, G.W. & Cochran, W.G. (1980). Statistical Methods (7th edition). Iowa State University Press, Ames, Iowa.

Appendix

Programs:

RISKNPAR.GEN and RISKNPAR.XLS implement the nonparametric approach of Chapter 3 in Genstat and @Risk, respectively.

RISKPAR.GEN implements the parametric method of Chapter 4 in Genstat.

Data files:

personen.lis: respondents (personal number, age, weight)

####_sto.lis: residue-labels (compound residuecode, label)

####_geh.lis: positive residue concentrations (productcode, concentration)

####_nge.lis: total number of detects and non-detects (x, productcode, n)

####_con.lis: consumption data (personal number, x, x, compound productcode, x)

####_prd.lis: product-labels (x, productcode, label)

####_pro.lis; same as prd file, but with an additional column indicating whether the pesticide is allowed (1) or not (0) for each product (this file is used in RISKPAR.GEN)

Replace '####' by code for residue (CHLR = Chlorothalonil, IPRO = Iprodione, PARA = Parathion, PIRI = Pirimicarb, TOLC = Tolclofos-methyl)

####_prd.lis

Each product is characterized by a product code built hierarchically from 5 numbers:

1 - productfile number

2 - productgroup number

3 - productsubgroup number

4 - productnumber

5 - productquality number

example:

productfile

1 food commodity

productgroup

1 7 pulse, seed, pits and nuts

2 8 vegetables, potatoes, beet and turnip

3 9 fruit

productsubgroup

1 7 1 pulse

1 8 1 leaf, stem and stalk vegetables

1 8 2 cabbages species

1	8	4 potatoes, carrots, turnip
1	9	2 apple species
1	9	4 berries

product (incl. productquality)

1	7	1	10	1 bean (scarlet runner, green bean, string bean)
1	8	1	2	1 endive
1	8	1	5	1 cabbage lettuce, bindsla
1	8	1	15	2 spinach
1	8	2	2	1 cauliflower
1	8	4	1	1 potato
1	8	4	3	1 beet
1	9	2	1	1 apple
1	9	4	3	1 strawberry

####_sto.lis

Each residue is characterized by a code built hierarchically from 3 numbers

1 – residue group number

2 – residue subgroup number

3 - residue number

example

residue group

11 bactericides and fungicides

21 elements and organometals

residue subgroup

11	5 dicarboximides
21	3 other anorganic compounds

residues

11	5	1 Iprodione (=glycofeen)
21	3	5 nitrate