A Background Document for the Session:

Statistical Methods for Use of Composite Data in Acute Dietary Exposure Assessment

of the May 26, 1999 Meeting of the FIFRA Scientific Advisory Panel

Executive Summary

The Environmental Protection Agency's (EPA) Office of Pesticide Programs (OPP) has identified a statistical methodology for applying existing information from the U.S. Department of Agriculture's (USDA) Pesticide Data Program (PDP) report to risk assessments of the acute exposure to pesticide residues in food. This methodology consists of extrapolating from data on pesticide residues in composite samples of fruits and vegetables to residue levels in single units of fruits and vegetables. Given the composite sample mean **($\bar{x}$),** the composite sample variance **(S^2),** and the number of units in each composite sample, it is possible to estimate the mean and variance **(μ** and **σ^2)** of the pesticide residues present on single units of fruits and vegetables. These parameters can then be applied to generate information on the level of residue in fruits and vegetables. This information can then be incorporated into a probabilistic exposure estimation model, such as the Monte-Carlo method, in order to estimate exposure to pesticide residues in foods and the risk attendant to that exposure. This methodology has a higher degree of accuracy when more than 30 composite samples have detectable residues. The statistical procedure outlined in this paper estimates the distribution parameters that describe residues of chemicals on single units of fruits and vegetables.

Other organizations have developed similar methodologies for extrapolating from residue levels in composite samples to residue levels in single units. These organizations include Sielken Inc. and Novigen Sciences, Inc. Because the methods developed by these two organizations originate from the same fundamental assumption that residues on individual serving sizes of fruits and vegetables follow a lognormal distribution, their results are similar to those of OPP.

OPP has recently started to apply the methodology described herein to estimate acute dietary exposure to pesticide residues in foods. OPP is asking the FIFRA Scientific Advisory Panel to answer specific questions regarding the methodology at the May 26, 1999 meeting of the Panel.

Introduction

As part of its process to assess the risks posed by pesticide substances, OPP estimates exposure to pesticide residues that results from consumption of foods that contain pesticide residues. Data for estimating dietary exposure to pesticide residues include the use of pesticide tolerances (i.e., the maximum concentration of a residue that can legally be present in an agricultural commodity or food); data obtained from field trials; and data obtained from monitoring studies. One of the principal sources of monitoring data is the USDA PDP annual survey. PDP collects and tests a variety of samples of foods commonly consumed in the United States. In many instances foods tested under the PDP have undergone some form of processing (e.g., cooking, washing, peeling, etc.). A primary advantage of using PDP data over tolerance levels or field trial data to assess dietary exposure is that PDP data represent actual concentrations of pesticide residues in the food supply close to the point of consumption, whereas the other data mentioned here typically do not. Hence, the availability of PDP data enables OPP to rely less on residue data that tend to overestimate the concentration of pesticide residues in foods, and make more realistic estimations of dietary exposure to pesticide residues. The PDP has been in operation since 1991, and data collected under this program are published annually. (More information on the PDP is available at: http://www.ams.usda.gov/science/pdp).

The process of assessing acute exposure to pesticide residues in food consists of multiplying the one-day food consumption data distribution by the concentration of pesticide residues on food, producing a distribution of chemical exposure for individuals in the United States.[7] PDP collects residue data not on individual units of food, (e.g., on a single apple), but rather on large (generally five pound) composite samples. After pureeing five pounds of apples, PDP proceeds to measure the residue content in that five-pound composite mass. Residue levels measured in the composite sample are likely to be close to the residues in food made by blending many individual units of the commodity (e.g., apple sauce). However, for many fruits and vegetables, people consume food in a form that is very different from the composite sample (e.g., it is a serving of a single unit of the commodity, such as a fresh apple). The residues on any given component may be higher or lower than the average residue level measured in the composite. Therefore, when a person eats a discrete unit of a commodity (e.g., a grapefruit), the acute food exposure assessment needs the values of residues on single units of the commodity, not the average residue in a composite sample.

The challenge to OPP has been to extrapolate from PDP composite data to provide single unit values for use in acute risk assessments. In statistical terms, given composite samples collected by the USDA, OPP is faced with the challenge of estimating the parameters that describe the original population of residue concentrations in servings of fruits or vegetables. Specifically, the problem is to estimate the population mean **(μ)** and the population variance **(σ^2)** from a set of composite samples where only the composite sample mean **($\bar{x}$)**, the composite sample variance **(S^2)**, and the number of units in each composite is known. With the estimation of the population parameters (μ and σ^2) and assuming that the distribution of residues in fruits and vegetables follows a lognormal distribution (as established in previous goodness-of-fit studies),[2, 4]

the function that describes chemical residues on fruits/vegetables is adequately established and ready for application into one of the components of the Monte-Carlo model for the acute risk assessment.

In order to apply the proposed methodology in connection with its related Monte-Carlo analysis, the PDP data are broken into three distinctive groups.

(1) **Zero residues:** Fruits and vegetables that have not received pesticide treatment at all. Calculated as the total crops minus the maximum percentage of crops treated (%CT) or 100% - maximum %CT.

(2) **Non-Detects:** Proportion of crops that, despite receiving treatment, have no detectable residues. Residue values for the composite sample of this group are between zero and the limit of detection (LOD).

(3) **Detects:** Group of composite samples that show residues in the PDP report.

To maintain the premise of lognormality, additional assumptions call for the separation of the PDP data representing non-treated fruits. Because there is a proportion of crops that has not been treated, it is assumed that this proportion has no residues (zero residues). Also, a group of composites exists that, despite receiving treatment, has values under the limit of detection (LOD). These are called non-detect values, and are so close to zero that lab instrumentation cannot detect any residue. Because these values are between zero and the LOD and nothing is known about them, these cases are characterized by other methods.

The methodology described in this paper deals only with group (3), or 'Detects,' and will estimate the parameters of the distribution for this group. All composites used for the parameter estimations have a definite, known value higher than or equal to the LOD; however, individual values within the composites may be zeros. All values are expressed in parts per million (ppm).

Questions for the Scientific Advisory Panel

(1) Measurements of many natural processes may be described by typical statistical distributions, e.g., normal, lognormal, etc. In previous data-fit studies, data on concentration of residues on fruits and vegetables have been fitted to a lognormal distribution. The lognormality of residues has been established as a fundamental assumption in the decomposition procedure. **Please comment on the assumption of lognormality.**

(2) The application of OPP's decomposition methodology calls for at least 30 "detects". This is done to assure that there is enough representation in the sample and that the extrapolation will cover the width of the distribution of single units.

Although 30 detects is a practical rule for the application of the procedure, please comment on the consideration of other numbers as a practical rule of application.

3. The standard deviation within a composite cannot be greater than the standard deviation of the population of individual residues. **Are there any circumstances when this statement is not true? If so, what are these circumstances?**

4. OPP acknowledges that the collection of composite samples in the PDP protocol is not purely random; therefore, the decomposition procedure will produce an overestimation of the standard deviation of the lognormal distributions of residues on fruits and vegetables. Moreover, the overestimation of the standard deviation is accentuated to the degree that the collection of composite samples departs from pure randomness. The consequence of overestimating the standard deviation is that the high end of the estimates of residues in single units may exceed what occurs in reality. **What criteria (if any) should be used to establish an upper-bound on the amount of residue projected in a single unit to address the potential for overestimation of the standard deviation?**

5. OPP's methodology is sensitive to the number (N) of single units/servings of a commodity estimated to be in a composite sample. **Please comment on how to estimate that number for different commodities.** (Consider how to handle fruits for which a single unit is typically only a part of a unit of a commodity (e.g., a melon) or many different units (e.g., grapes) even though the single unit is smaller than the typical composite sample.)

6. The decompositing procedure estimates the number of units in a PDP composite by dividing the weight of the composite by an average weight of an individual unit. The number of individual units in a composite may vary, depending upon the weight of each component unit. **Will such differences in the number of individual units introduce substantial uncertainty?**

Methodology

After an extensive research of the existing statistical literature, EPA has found a procedure that evaluates the problem at hand. Dixon and Massey[(1)] state that "if a collection of means **($\bar{x}$)** of samples of size N are available, the **$\bar{x}$** 's may be used to estimate the mean **(μ)** and the variance **(σ^2)** of the original population." The elements necessary to estimate the population's parameters **(μ** and **σ^2)** are:

(1) The mean ($\bar{x}_i$) of each sample collected (value of each "detect" composite sample in the PDP data),

(2) The number of samples (n) collected (the number of samples of the chemical-fruit pair collected in the PDP data), and

(3) The number of units (N) within each sample (the amount of servings in each PDP composite).

The PDP data provide the mean of each composite sample collected ($\bar{x}_i$). In previous studies conducted by Novartis on peaches,[2] it has been shown that the residue detected on the composite equals the average residue of the component units. PDP data also contain the number of samples collected (n); however, it does not explicitly state the number of units (N) there are in each sample (e.g. how many apples are in a composite sample). Nevertheless, it is possible to estimate the number of units or serving sizes that goes into a composite sample. According to the 1996 PDP Annual Summary,[3] under Section II, *Sampling Protocol*, "the sample size was approximately 5 pounds for fresh products, 3 pounds for canned and frozen product, and 1 quart for liquid juice for each applicable testing facility."

Knowing the average weight of an individual unit, it is easy to estimate the number of units in a composite sample. For example, in the Ministry of Agriculture, Fisheries & Food (MAFF) study,[4] the average weight of an apple serving is estimated at 142 grams. The PDP protocol indicates that 5 pounds of apples form a composite sample; therefore, it is possible to conclude that approximately 16 units are in a composite sample of apples (5 pounds equal about 2,267 grams). In the case that a composite sample contains different numbers of individual units (i.e., one composite sample of apples may contain 14 individual units, while another composite sample may contain 16 individual units), the composite sample with the highest number of individual units should be used in order to achieve a wider range of variability. Using this information, one can proceed to estimate the parameters of the lognormal distribution of residues.

The parameter estimation consists of two steps:

(1) Estimation of the mean and variance of the sample of composite values, and

(2) The adjustment of the sample variance that corrects for the units making up the composite and that extrapolates the composite samples to the original population of residues.

Step One: Estimation of Mean and Variance from the Sample of Composites

Parameter estimation can be done by transforming the original data into logarithms and then calculating the sample mean and variance of the logarithmic data. Plugging these values into the transformation equations for the lognormal (shown below), the estimated parameters can be

obtained.

To transform the log data into the parameters of the original lognormal distribution the following equations are used:

$$\bar{X}_{est} = \exp\left(\bar{X}_{log} + \tfrac{1}{2} S^2_{log}\right)$$

$$S^2_{est} = \exp\left(2\,\bar{X}_{log} + 2\,S^2_{log}\right) - \exp\left(2\,\bar{X}_{log} + S^2_{log}\right)$$

Where $\bar{X}_{log}$ and S^2_{log} are the mean and variance of the logarithmic data, and $\bar{X}_{est}$ and S^2_{est} are estimates of mean and variance of the composite sample.

Step Two: Corrections for the Variance

According to Dixon and Massey,[1] the best estimator of μ, the mean of the population, is the mean of the composite sample or the value $\bar{X}_{est}$ found in step one. The best estimator of the variance of the population (σ^2) is the variance S^2_{est} multiplied by the number of units in each sample (N = 16) or:

$$\mu = \bar{X}_{est}$$

$$\sigma^2 = (S^2_{est})\,(N)$$

To estimate the standard deviation, the correction for the equation of the variance can be expressed as the following:

$$\sigma = S_{est}\sqrt{N}$$

For the validity of these equations it is assumed that, when forming the composite samples, units are drawn at random and as independent events. For example, the sixteen servings of apples going into a composite have an equal chance of coming from anywhere in the universe of residues, and values ranging from zero to the maximum could coexist in a composite. However, PDP builds the composites from apples from the same box located at warehouses or collection points. Despite the fact that the box is selected at random, the composite sample formed this way does not comply with the assumption of total randomness because apples in the same box are likely to have had the same pesticide treatment and history. Thus, while the residues on individual apples may differ, the variation is likely to be smaller than variation among apples that have had different treatment histories. Although this fact has no impact on the estimation of the mean value, it produces an overestimation of the variance of the population of residues that

maintains the conservative nature of the risk assessment. Therefore, the method overestimates the variance of the residues in fruits and vegetables in the degree that the collection of samples departs from pure randomness.

Confidence Intervals for the Estimated Parameters

With the estimation of the mean and the corrected variance of the population of individual values, the lognormal distribution describing the individual residues is defined and any individual value of residues can be generated from it. Moreover, it is possible to calculate confidence intervals for the estimated parameters. Confidence intervals give both an idea of the actual, numerical value the parameters may have, and also an indication of the confidence level, on the basis of the sample, that a correct indication of the possible numerical value of the parameter has been given. Once the extreme values of the interval are calculated, the confidence interval bracketed by the calculated extremes containing the unknown value of the parameters is 100 (1-**α**)%. The term **α** signifies the probability of not having the estimated parameter in the confidence interval.

For the population mean **(μ)** the confidence interval with a certainty of 100(1- α)% is given by the following inequality:

$$\bar{\mathbf{X}} - t_{1-\alpha/2}\,\mathbf{S_{est}} / \sqrt{\mathbf{n}} \;<\; \mu \;<\; \bar{\mathbf{X}} + t_{1-\alpha/2}\,\mathbf{S_{est}} / \sqrt{\mathbf{n}}$$

These limits provide **100(1 - α)%** assurance that the population mean **(**μ**)** is contained in the interval. The portion $t_{1-\alpha/2}$ is the 100(1 - α/2) percentile of the t-distribution with n -1 degrees of freedom. A table for the t-distribution is found in Larson.[5] The value **n** is the number of samples, and $\mathbf{S_{est}}$ is the standard deviation of that sample.

Similarly, a confidence interval for the population's standard deviation (**σ**) is given by the inequality:

$$\mathbf{S_{est}}\sqrt{\mathbf{N}} / \left({}^{2}_{1-\alpha/2}/\mathrm{df} \right)^{1/2} \;<\; \sigma \;<\; \mathbf{S_{est}}\sqrt{\mathbf{N}} / \left({}^{2}_{\alpha/2}/\mathrm{df} \right)^{1/2}$$

As before, these limits contain the population's standard deviation **(σ)** with 100(1 - **α**)% assurance that the standard deviation will be held in. The portion ${}^{2}_{1-\alpha/2}/\mathrm{df}$ and ${}^{2}_{\alpha/2}/\mathrm{df}$ are the 100(1 - **α/2**) and the 100(**α/2**) percentiles of the ${}^{2}/\mathrm{df}$ distribution with n -1 degrees of freedom. A table with the values of this distribution is available in Dixon and Massey.[1] The value N is the number of units in each composite sample, and $\mathbf{S_{est}}$ is the estimated sample standard deviation.

Demonstration of how the parameter estimation procedure works

For purposes of illustrating how this procedure operates, the EPA applied the statistical techniques described above to a real file containing data on carbaryl residues in 108 individual samples of U.S. apples[4] (see attachment 1). Assuming that the 108 samples represent the entire spectrum of carbaryl residues on apples, calculation of the mean value ($\bar{\mathbf{X}}$) and the standard deviation ($\mathbf{S_x}$) is as follows. Because of the assumption that these data constitute the entire spectrum of residues, these values can be considered, for the sake of the example, the population's parameters μ and $\boldsymbol{\sigma}$, calculated as μ **= 1.41 and σ = 0.71** (table 1).

Table 1: Results of the calculations on the 108 residues of carbaryl in apples (considered as population parameters)

Sample Statistics	
Average Concentration	1.411667
Standard Deviation	0.707506

A goodness-of-fit test shows that the data points follow a lognormal distribution. An examination of the methodology consists of arriving independently at a confidence interval containing these parameters having only samples that simulate the conditions of composite sampling. This implies that the composite sample mean is known, but the sample's standard deviation is unknown. To simulate composite samples, ten random samples are picked (n = 10), with five apples in each sample (N =5). This way, ten $\bar{\mathbf{x}}_\mathbf{I}$ are obtained as shown in table 2 below.

Table 2: Composite Sampling Simulation

	No. 1	**No. 2**	**No. 3**	**No. 4**	**No. 5**	**No. 6**	**No. 7**	**No. 8**	**No. 9**	**No. 10**
	1.06	2.31	1.58	1.78	1.18	1.63	0.44	1.56	1.78	0.63
	1.73	1.58	1.07	0.26	1.32	0.63	1.28	1.49	1.49	0.8
	1.1	0.63	0.85	0.81	1.82	1.3	2.3	1.2	0.63	0.26
	0.99	0.3	3.89	0.81	1.25	0.31	0.96	0.69	2.72	1.68
	1	0.8	0.65	1.52	1.61	2.43	1.29	3.43	1.3	3.89
Average	**1.176**	**1.124**	**1.608**	**1.036**	**1.436**	**1.26**	**1.254**	**1.674**	**1.584**	**1.452**

This set of ten composite samples is transformed into logarithms and the mean and

variance are calculated as shown in table 3, resulting in:

Log-Mean = 0.2957, and
Log-Variance = 0.0271

Table 3: Transformation of the composite sample into logarithmic data

	Composite Samples (n=10, N=5)	Logarithmic Data
	1.176	0.16212
	1.124	0.11689
	1.608	0.47499
	1.036	0.03537
	1.436	0.36186
	1.26	0.23111
	1.254	0.22634
	1.674	0.51522
	1.584	0.45995
	1.452	0.37294
Mean	**1.3604**	**0.29568**
Variance	**0.04892**	**0.02714**

Applying the transformation equations results in:

$$\bar{X} = \exp (0.2957 + 0.0271/2)$$

$$\bar{X} = 1.362$$

$$S_{est}^{2} = \exp [2(0.2957) + 2(0.0271)] - \exp [2(0.2957) + 0.0271]$$

$$S_{est}^{2} = 0.051$$

or

$$S_{est} = 0.225$$

$$\bar{X} = 1.36 \text{ and } S_{est} = 0.225$$

The value $\bar{X} = 1.362$ is the estimation of the mean value **(μ)** of the population; the 90% confidence interval ($\alpha = 0.1$) for this sample is:

$$\bar{X} - t_{1-\alpha/2} S_{est} / \sqrt{n} < \mu < \bar{X} + t_{1-\alpha/2} S_{est} / \sqrt{n}$$

Where:

$\bar{X} = 1.362$,

$t_{1-\alpha/2} = 1.833$ is the ninety-five percentile with (10 - 1) = 9 degrees of freedom of the t-distribution,

$S_{est} = 0.225$, and

$n = 10$.

$$1.362 - (1.833)(0.225)/3.16 < \mu < 1.362 + (1.833)(0.22)/3.16$$

$$1.362 - 0.13 < \mu < 1.362 + 0.13$$

$$1.23 < \mu < 1.49$$

Thus, one can say with 90% confidence that the mean value of the population lies between the extremes of the above interval. Previous calculation of the population mean produced a value of $\mu = 1.41$ which corroborates that the estimated mean is contained in the calculated interval.

The calculations for the correction of the standard deviation are:

$$\sigma = S_{est} \sqrt{N}$$

or, $\sigma = (0.22)(2.24) = 0.49$

A confident interval for σ is given by:

$$S_x \sqrt{N} / ({}^2_{1-\alpha/2} /df)^{1/2} < \sigma < S_x \sqrt{N} / ({}^2_{\alpha/2} /df)^{1/2}$$

Where:

$S_x = 0.22$,

${}^2_{1-\alpha/2}/df = 1.880$ is the ninety-five percentile of the ${}^2/df$ distribution function with (10 - 1) = 9 degrees of freedom,

$^2_{\alpha/2}$ /df = **0.369** is the five percentile of the ^{2} /df distribution function with (10 - 1) = 9 degrees of freedom, and

N = 5

$$0.22(2.24)/\sqrt{1.88} < \sigma < 0.22(2.24)/\sqrt{0.369}$$

$$0.359 < \sigma < 0.811$$

Therefore, one can say with 90% confidence that the standard deviation of the population lies between the extremes of the above interval. Calculation of the population's standard deviation produces a value of σ = 0.71 that corroborates the above statement.

The same methodology may be applied to any set of composite samples and intervals of confidence drawn to cover different percentage of assurance (α = 0.05, α = 0.1, etc.) for the contention of the parameter in the interval's limits.

However, the objective of this exercise is to provide a set of numbers comparable to the original set of numbers shown in attachment 1. To achieve this, with the estimated parameters (mean 1.36 and standard deviation 0.49) random values of a lognormal distribution are generated using, in this case, the commercially available Crystal Ball™ software. Attachment 2 shows 108 randomly generated values (Crystal Ball™: version 4.0, Monte Carlo sampling method: Latin Hypercube, Initial Seed Value: 1). By comparing these two value sets, the observer can conclude that, despite the small sample which introduces sampling error, the two data sets are very similar.

Assessment of the Overestimation Produced by the Methodology

Several computer generated data sets have confirmed that when restrictions are applied to the way composite samples are collected (e.g., when the individual commodities in a composite sample are not included randomly), the decomposition methodology produces an overestimation of the standard deviation of the lognormal distribution of residues. The data also show that the more severe the restrictions are (less variability within the composite), the greater the overestimation of the standard deviation. However, because of a lack of appropriate data, OPP cannot state how much the decomposition procedure may overstate actual residue levels in single unit samples.

OPP is interested in better characterizing the level (or range) of overestimation that typically occurs when residue levels in single unit samples are estimated from a distribution of residue levels in composite samples. In order to evaluate the difference between actual residue levels and estimated residue levels in individual units, the Agency would need the following data for a number of composite samples of a commodity (ideally at least 30 composite samples): the levels of residues in the composite and the levels of residues in each of the individual items that comprise the composite sample. For example, if a standard composite sample typically contains 16 apples, the data should be generated by cutting each apple in half and putting one half of each apple into a composite and measuring the average residue in the composite. The residue levels in

each of the other 16 halves would also be determined. The Agency could then compare its estimates of the single unit samples using the composites with the actual measurements in the individual apples. OPP anticipates that the results may be quite different, depending on the variability in the pesticide use patterns for a particular commodity. Therefore, OPP invites submission of data on a variety of commodities to allow the assessment of this issue.

Acknowledgment of Other Imputation Procedures

In addition to the decompositing method developed by OPP, other organizations have developed methodologies that extrapolate the composite samples to single units (decomposition). These methodologies include the "Maximum Likelihood Imputation Procedure for Imputing Single-Serving Residue Concentration Distributions from Composite Samples," developed by Sielken Inc., and "NSI Multi-Distribution Imputation Procedure," developed by Novigen Sciences, Inc.

The three procedures developed by OPP, Sielken Inc., and Novigen Sciences, Inc. originate from the fundamental assumption that residues on individual serving sizes of fruits and vegetables follow a lognormal distribution. Because the objective of the decomposition is to complement the Dietary Exposure Evaluation Model (DEEM) software, the result of the decomposition is a file of values representing residues on single units.

From this fundamental assumption, Sielken developed a computer program called MAXLIP that consists of two steps. In the first step, MAXLIP finds an approximation to the individual unit residue concentration distribution using a family of statistical distributions (the lognormal distributions or mixtures of lognormal distributions). Then, in the second step, the approximate theoretical distribution is used to create a distribution of sample individual unit residue concentration values that is closer to, and less dependent on, the theoretical characteristics of the lognormal distribution. MAXLIP repeats the maximum likelihood procedure thousands of times and creates a file with the estimated values of the residues on single units.

The Novigen methodology provides two methods for estimating the mean and standard deviation of the lognormal. One uses the Central Limit Theorem, and the other uses a fixed coefficient of variation (CV = 1), based on average CV from field trial residue data. Based on these two methods, Novigen developed the computer program KEVIN.EXE that calculates residues for each composite, thus creating a file of residues in individual unit samples.

OPP's methodology assumes that residues of a given pesticide on fruits and vegetables form a single universal lognormal distribution, due to the huge amount of single units produced in the United States. Composite values coming from this unique distribution provide a variability reflective of the many existing conditions. The method calculates the mean value and the adjusted standard deviation of the assumed lognormal distribution; once those parameters are identified, a

data file of residues corresponding to that lognormal distribution are generated.

Conclusion

With the estimates of the mean and variance and assuming that residues in fruits and vegetables follow a lognormal distribution, the distribution of residues of a specific chemical can be estimated for individual units by just knowing the $\bar{x}_{I's}$ of a sample set and the number of units within each composite sample. Identifying the distribution form and its parameters establishes the distribution itself; thus, this calculation allows the use of the PDP data to be extended to the single unit analysis used in the acute risk assessment. Once the mean and variance have been estimated, and knowing that the distribution is lognormal, the data can be entered into a software program, such as Crystal Ball™, in order to generate as many data points as necessary.

This methodology will increase its accuracy to the degree that more samples are collected from the same population. Usually, the estimation is better for more than 30 samples (n =30).[6] In practical terms, the accuracy will deteriorate if the number of units in the composite sample (N) is much larger than the number of samples (n), i.e., **N > > n**, and the number of samples is small, for example, 7 samples of 100 apples each. However, none of these hypothetical conditions is present in the PDP data collection.

OPP recognizes that the decomposition procedure described in this paper produces an overestimation of the standard deviation of the lognormal distribution of residues on fruits and vegetables, and invites the submission of data that will allow the Agency to assess the difference between actual and estimated residue levels.

References:

[1] Dixon, W. and F. Massey, 1957. Introduction to Statistical Analysis. McGraw-Hill, New York, NY.

[2] Sielken, Robert L. Jr., 1998. "Maximum Likelihood Imputation Procedure for Imputing Single-Serving Residue Concentration Distributions from Composite Samples." Presented to EPA/OPP, Washington, D.C., December 10, 1998.

[3] USDA, 1996. "PDP Annual Summary Calendar Year 1996."

[4] Ministry of Agriculture, Fisheries & Food (MAFF), 1997. "Unit to Unit Variation of Pesticide Residues in Fruit and Vegetables." March 14, 1997.

[5] Larson, H., 1974. Introduction to Probability Theory and Statistical Inference. John Wiley &

Sons, New York, NY.

(6) Walpole, Ronald E., 1982. Introduction to Statistics. Macmillan Publishing Co., Inc., New York, NY.

(7) EPA, 1999, "A User's Guide to Available OPP Information on Assessing Dietary (Food) Exposure to Pesticides," Notice of Availability, Federal Register, January 4, 1999.

Attachment 1: Carbaryl residue data in 108 samples of U.S. apples (Harpenden; Apple 4995, Ion Trap Mass Spectrometer)**. Values are sorted descending.**

Sample #	*Concentration (mg/kg)*	Sample #	*Concentration (mg/kg)*	Sample #	*Concentration (mg/kg)*
1	***3.89***	**41**	***1.49***	**81**	***0.94***
2	***3.43***	**42**	***1.49***	**82**	***0.93***
3	***2.97***	**43**	***1.46***	**83**	***0.92***
4	***2.82***	**44**	***1.45***	**84**	***0.89***
5	***2.72***	**45**	***1.45***	**85**	***0.85***
6	***2.67***	**46**	***1.44***	**86**	***0.81***
7	***2.63***	**47**	***1.39***	**87**	***0.80***
8	***2.60***	**48**	***1.38***	**88**	***0.80***
9	***2.56***	**49**	***1.35***	**89**	***0.80***
10	***2.56***	**50**	***1.32***	**90**	***0.80***
11	***2.51***	**51**	***1.32***	**91**	***0.79***
12	***2.50***	**52**	***1.32***	**92**	***0.78***
13	***2.43***	**53**	***1.30***	**93**	***0.77***
14	***2.41***	**54**	***1.30***	**94**	***0.71***
15	***2.40***	**55**	***1.30***	**95**	***0.70***
16	***2.32***	**56**	***1.29***	**96**	***0.69***
17	***2.31***	**57**	***1.28***	**97**	***0.69***
18	***2.30***	**58**	***1.27***	**98**	***0.64***
19	***2.08***	**59**	***1.25***	**99**	***0.63***
2	***2.07***	**60**	***1.20***	**100**	***0.63***
21	***1.98***	**61**	***1.19***	**101**	***0.60***
22	***1.92***	**62**	***1.18***	**102**	***0.55***
23	***1.90***	**63**	***1.18***	**103**	***0.45***

Sample #	*Concentration (mg/kg)*	Sample #	*Concentration (mg/kg)*	Sample #	*Concentration (mg/kg)*
24	***1.81***	**64**	***1.12***	**104**	***0.34***
25	***1.78***	**65**	***1.10***	**105**	***0.31***
26	***1.74***	**66**	***1.09***	**106**	***0.30***
27	***1.73***	**67**	***1.09***	**107**	***0.26***
28	***1.68***	**68**	***1.08***	**108**	***0.25***
29	***1.65***	**69**	***1.08***		
30	***1.63***	**70**	***1.07***		
31	***1.63***	**71**	***1.06***		
32	***1.61***	**72**	***1.00***		
33	***1.61***	**73**	***1.00***		
34	***1.59***	**74**	***1.00***		
35	***1.58***	**75**	***0.99***		
36	***1.58***	**76**	***0.99***		
37	***1.56***	**77**	***0.99***		
38	***1.52***	**78**	***0.97***		
39	***1.51***	**79**	***0.96***		
40	***1.49***	**80**	***0.94***		
MEAN = 1.411666667					
STANDARD DEVIATION = 0.707506468					

Attachment 2: Randomly Generated Residue data in 108 samples (Crystal Ball™: version 4.0, Monte Carlo sampling method: Latin Hypercube, Initial Seed Value: 1) **Values are sorted in descending order.**

Sample #	*Concentration (mg/kg)*	Sample #	*Concentration (mg/kg)*	Sample #	*Concentration (mg/kg)*
1	***3.97***	**41**	***1.42***	**81**	***1.03***
2	***3.29***	**42**	***1.41***	**82**	***1.02***
3	***2.76***	**43**	***1.40***	**83**	***0.99***
4	***2.67***	**44**	***1.39***	**84**	***0.96***
5	***2.27***	**45**	***1.38***	**85**	***0.96***
6	***2.22***	**46**	***1.36***	**86**	***0.95***
7	***2.21***	**47**	***1.34***	**87**	***0.95***
8	***2.17***	**48**	***1.34***	**88**	***0.93***
9	***2.10***	**49**	***1.32***	**89**	***0.91***
10	***2.06***	**50**	***1.32***	**90**	***0.91***
11	***2.02***	**51**	***1.30***	**91**	***0.90***
12	***2.00***	**52**	***1.29***	**92**	***0.89***
13	***1.95***	**53**	***1.29***	**93**	***0.89***
14	***1.94***	**54**	***1.28***	**94**	***0.88***
15	***1.81***	**55**	***1.28***	**95**	***0.85***
16	***1.74***	**56**	***1.27***	**96**	***0.85***
17	***1.73***	**57**	***1.22***	**97**	***0.81***
18	***1.72***	**58**	***1.22***	**98**	***0.80***
19	***1.69***	**59**	***1.21***	**99**	***0.79***
20	***1.67***	**60**	***1.21***	**100**	***0.78***
21	***1.67***	**61**	***1.20***	**101**	***0.77***
22	***1.66***	**62**	***1.20***	**102**	***0.75***

Sample #	*Concentration (mg/kg)*	Sample #	*Concentration (mg/kg)*	Sample #	*Concentration (mg/kg)*
23	***1.65***	**63**	***1.20***	**103**	***0.73***
24	***1.64***	**64**	***1.19***	**104**	***0.73***
25	***1.62***	**65**	***1.19***	**105**	***0.70***
26	***1.61***	**66**	***1.17***	**106**	***0.68***
27	***1.60***	**67**	***1.17***	**107**	***0.65***
28	***1.59***	**68**	***1.17***	**108**	***0.50***
29	***1.58***	**69**	***1.16***		
30	***1.56***	**70**	***1.15***		
31	***1.56***	**71**	***1.14***		
32	***1.53***	**72**	***1.13***		
33	***1.53***	**73**	***1.12***		
34	***1.53***	**74**	***1.12***		
35	***1.52***	**75**	***1.10***		
36	***1.51***	**76**	***1.09***		
37	***1.50***	**77**	***1.09***		
38	***1.43***	**78**	***1.07***		
39	***1.43***	**79**	***1.06***		
40	***1.42***	**80**	***1.05***		
MEAN = 1.36728					
STANDARD DEVIATION = 0.53405					