Introduction - Sampling Working Group

As mentioned in the executive summary on sampling, an understanding of the sampling and measurement procedures is necessary for obtaining confidence that the obtained results "represent" the intended population or fulfill a study's purpose. The confidence of results can be undermined if care is not taken to control and minimize the variation of observed results due to sampling and measurement. To address this concern, the information below is presented as a foundation for and linkage to the two documents on measurement error (Enclosure A) and statistical process control sampling (Enclosure B).

Sampling and Measurement

An important reason, and the one that is of interest for this Committee, for analyzing samples in the first place is to characterize some aspect of the distribution of the "true level", x, or most generally, to determine the distribution of x, within some well-defined population of product that the analyzed samples are "representing." The values of x could refer to levels or densities of some measurand or could refer to whether or not a pathogen is present in a sampled material. The results are a collection, $\{y_j, j = 1, \ldots n\}$ where n is the number of samples (here assumed randomly drawn for some population, with equal probability of selection). Thus values of y refer to the measured result, either a measurement of level or density of some measurand, or whether or not the pathogen was found. The value of y thus represents the "known" evidence, from which an inference is made regarding the possible values of x. In the interferential process there is always uncertainty associated with any conclusion or characterization made about possible value of x.

Mathematically this uncertainty can be represented by a "likelihood" function. This function can be derived in stages. First consider the probabilities of possible values of y for hypothetical values of x, $g(y|x)$. This is a function of the true value of x. However, x is not known, but rather y is known. The values of x, being unknown, are (next) assumed to occur with some probability density which can be labeled, f(x). With this supposition, the (full) probability relationship between y and x can be written down mathematically. To distinguish the case of y being known and x being unknown (from the case of x being known) the phrase "probability of y" is not used, but rather the phrase used is the "likelihood of y." More specifically, if the density of the distribution of x is f(x), then the likelihood (L) of obtaining a value of y can be expressed as a joint probability integral equation:

$$L(y) = \int g(y|x)f(x)dx \qquad (1)$$

This equation includes results reported as non-detects, ND, as a possible value. That is,

$$L(ND) = \int g(ND|x)f(x)dx \qquad (2)$$

where g(ND|x) is the probability of ND (of getting a non-detect, or a false negative) given a true value of x in the sample. An estimate of f(x) can be derived from the above integral equation, assuming g(y|x) is known. If f(x) is of known (or assumed to be a specific) mathematical form, parameterized with parameter vector, θ, of (often) unknown values, then using maximum likelihood (MLE) estimation or method of moments (MOM), estimates of values of θ can be obtained.

Often, the forms of f(x) and even g(y|x) will not be known, but their first two moments (mean and variance) can be estimated and be considered sufficient for many purposes. An example of this is with statistical process control (SPC), discussed in Enclosure B (SPC document), where SPC procedures depend upon specifying the mean and variance of the process. If it is assumed that the relationship between the expected value and variance of y given x and x is known, then the mean and variance of the distribution of x can be obtained. The relationship for the means and the variances are:

$$E(y) = E_f(E(y|x)) \quad (4)$$

$$Var(y) = var_f(E(y|x)) + E_f(var(y|x)) \quad (5)$$

where E_f and var_f refer to the expected value and variance of the distribution with density function f, and E and var, without subscripts, refer to expected value and variance of distribution g. The terms on the left are determined directly from the collection of $\{y_j, j = 1, \ldots n\}$of sample results; the terms E(y|x) and var(y|x) are assumed known functions of x, so that the above equations can be used to solve for $E_f(x)$ and $var_f(x)$.

A simple example is to assume that var(y|x) is some linear function of x: var(x) = ax + b, where a and b are constants. For some methods, such as methods of measuring densities of chemical residues, the coefficient of variability (CV) is assumed to be equal to 100(a + b/x), when x is the true level of some analyte, so that the variance, var(y|x), would be $(ax+b)^2$. Assuming that E(y|x) = x – that is, the method is unbiased - the above equations become:

$$E(y) = E_f(x) \quad (4a)$$

$$Var(y) = var_f(x) + E_f(ax+b) \text{ , or in the second case,} \quad (5a)$$

$$Var(y) =[1 + a^2]\, var_f(x) + [aE_f(x) + b]^2 \quad (5b)$$

If E(y) and Var(y) can be estimated from *a priori* information, for example, from inter- or intra-laboratory studies, then $E_f(x)$ and $var_f(x)$ can be estimated by solving the above equations.

Often there is a need for imputation or assigning a value of y when the imputed value is a non-detect value (ND). A standard procedure for imputation is to impute ½ the limit of detection (LOD) (EPA, 2000), and then compute the average and standard deviation using the imputed values for ND. A justification of this imputation procedure could be based on the

"principle of indifference,"[1] which here would invoke an assumption that the values of y that could have been measured would be uniformly distributed between 0 and L (where L = LOD). In other words, if it is thought that y represents an estimate of x on a sample, then the "best" estimate of x given that y is below the LOD = L, by the principle of indifference, is L/2. This is a confusing assumption and its very premise leads to contradictions, as is well known; for example, by the same "principle of indifference" applied to the square root of the true level, $x^{1/2}$, the imputed estimate would be $L^{1/2}/2$, so that for x, the imputed estimate would be the square of this value, specifically, L/4. Ideally if the true distribution were known (or assumed) then values for ND results could be derived using statistical estimation procedures. Based on assumptions for the distribution, procedures for imputation of results reported below the LOD have been proposed (Cohen, 1959; Persson and Rootzen, 1977; Singh and Nocerino, 2002). In any case, at least with chemical measurements, the LOD/2 imputation is commonly used (EPA, 2000) and would permit the above calculations to proceed.

Importance for sampling

It might be (as is often the case) that the percentage of the variance component (of the total variance) due to measurement is small relative to the variance component due to sampling variation. However, even in this situation, the variance of individual results can be of such magnitude to affect significantly the confidence that is associated with individual results. For a simple example, assume that the distribution of APC counts is lognormal, and that the mean of the $\log_{10}$ of the sample values, y, is 3 and the sample standard deviation is 1. Since we are assuming that the distribution of the $\log_{10}(y)$ is normal, a 95% probability interval would be approximately 1 ($\log_{10}$) to 5 ($\log_{10}$). Consequently, if there were a specification that "permitted" no more than 4.5 $\log_{10}$ on a sample[2], then based on the normal distribution for the logarithm of the APC counts, assuming a mean value of 3 $\log_{10}$ and a standard deviation of 1, there is a probability of 6.7% that a sample value would exceed 4.5 $\log_{10}$ (the z-score corresponding to the limit, 4.5 $\log_{10}$ is z =(4.5-3)/1 = 1.5, which has associated cumulative probability of 93.3%, so that probability of being greater than 4.5 is 6.7%).

For simplicity here, assume that the distribution of $\log_{10}(y)$, given a sample with a true level of x, such that the expected value of $\log_{10}(y)$ is $\log_{10}(x)$, and the standard deviation is 0.3 $\log_{10}$, independent of the value of x. From Equation 5a, $1^2 = \mathrm{Var}(\log_{10}(y)) = \mathrm{var}_f(\log_{10}(x)) + 0.3^2$, so that the population variance of $\log_{10}(x)$ is $1- 0.3^2 = 0.91$; and the standard deviation of $\log_{10}(x)$ is $(0.91)^{1/2} = 0.954$. Hence the 95% probability interval, symmetric about the mean of the $\log_{10}$ of the true levels for the population, is 1.13 to 4.87 $\log_{10}$ and there would be a 6.3% probability that a sample value would exceed 4.5 $\log_{10}$. The difference between the two

[1] Also referred to as the "principle of insufficient reason" developed in the 19th century, and later renamed 'priciple of indifference' by the economist John Maynard Keynes (http://en.wikipedia.org). It basically stipulates that lacking any other information one can assume equal probablites for a set of events. Where the events refer to values of continuous variables the principle leads to ambiguity as described within the text.

[2] In some situations, a specification would refer to the true level in a sample so that it would be necessary to know the measurement error to determine compliance. Some adjustment might be made then to account for measurement error.

intervals is not large (1 to 5 versus 1.13 to 4.87 $\log_{10}$). Now, if the measurement standard deviation were reduced by a factor of 2, to 0.15 $\log_{10}$, then the probability of a single result obtained on a randomly drawn sample being greater than 4.5 $\log_{10}$ would be about 6.0%, reduced from 6.3%, hardly a change at all.

On the other hand, the impression of the effect of reducing the standard deviation of the measurement error could be different when considering its impact on inferring a value for sample using single measurements. For a single measurement, a standard deviation of 0.3 $\log_{10}$ would imply that, the 95% confidence interval associated with that true sample value of $\log_{10}(x)$, would be $\log_{10}(y) - 0.588$, $\log_{10}(y) + 0.588$, a range of 1.176 $\log_{10}$, or a factor of about 15. If the true value for a sample was 4 $\log_{10}$, which is well below the specified limit amount of 4.5 $\log_{10}$, there would be about a 5% chance that a measured value would exceed 4.5, assuming a standard deviation of 0.3 $\log_{10}$ for the measured result. If the standard deviation were reduced by a factor of 2, then the range of the 95% confidence interval associated with a measured value would be $\log_{10}(y) - 0.294$, $\log_{10}(y) + 0.294$, a range of 0.588 $\log_{10}$, or a factor of about 3.9, a seemingly substantial reduction. The probability of a result being greater than 4.5 $\log_{10}$ given a true $\log_{10}$ value of 4 would be 0.043%, virtually zero, compared to the 5% when the measurement standard deviation is 0.3 $\log_{10}$. This could be considered a significant change.

Thus, overall, when considering the effect on sampling populations, reducing the measurement standard deviation from 0.30 to 0.15 does not amount to a significant change in the operating characteristic (OC) curve (which provides the probability of acceptable results given assumed true conditions (Juran, JM, 1951) when the results of the measurements are being used for assessing a distribution of levels within some population - in our example, the probability of failing was reduced from 6.3% to 6.0%, about a 5% reduction of the probability of obtaining failed samples. The effort needed to reduce the standard deviation by a factor of 2 would be at least 4 samples per analysis, and perhaps more, as discussed below. As shown by way of this example, it may not be worth the extra time and effort to increase the number of analyses per sample. However, when inferring a true value for a specific sample, perhaps in a legal setting, the reduction of the standard error of the mean might be significant, as illustrated by the above example.

In determining how many samples would be needed to reduce the standard error of the mean (compared to the standard deviation of a single result), the magnitude of the variance components associated with the sampling and measurements would need to be known. For example, very simply, the standard deviation may include significant day-to-day effects. In other words, samples analyzed on the same day would not be independent results, but rather would be correlated within the population of possible results that would be obtained for the sample if it were analyzed on different days with different reagents and so forth. This notion is expressed by identifying a parameter, δ, called the intra-day correlation, which is the proportion of the between-day variance to the total variance - that is, the sum of the between-and within-day variance. For n samples analyzed per day for m days, the variance of the mean would be

$$\sigma_m^2 = \sigma_0^2\delta/m + \sigma_0^2(1-\delta)/(mn) \qquad (6)$$

where the first term on the right side represents the contribution of the between-day variance component (sampling for m days), and the second term represents the contribution of the within-day variance component (for mn samples). For example, for a value of δ of 0.3, the mean of 56 samples, analyzed over 8 days, 7 samples per day has the same variance as that of the mean of 98 samples analyzed over 7 days, 14 samples per day. An intra-day correlation of 0.3 is large, but not unbelievable, particularly for microbiological measurements wherein "causes" of contamination or high levels of organisms could vary day- to-day by substantial amounts.

Assume that a result needs to be obtained daily for some quality assurance or control purpose and thus results are analyzed in one day, so that m = 1. A question might be: how many samples are needed (in one day) in order that the standard error (of the mean) is a fraction r of the standard deviation, σ_0, of a single result? From Equation 6, assuming $\delta < r^2$, the number of samples needed would be:

$$n = \frac{1-\delta}{r^2-\delta} \qquad (7)$$

Thus, for example, if $\delta = 0.1$ and r = ½, 6 samples per day would be needed to have a variance of the mean be ½ the variance of a single result.

Summary

For microbiological measurements, true levels of the measurand are often highly variable over time, so that in general, given resources for a fixed number of samples, more samples over time with less samples per day, and more days of sampling is preferable if the purpose of sampling is to examine trends or get a good profile of the distribution of the measurand over time. However, if decisions are to be made on sample results for a given day, to ensure that product is safe then more samples per day might be needed.

Composite sampling

To minimize costs, composite sampling can be considered, when k samples (for example in one day) are divided into m composites of n samples (so that k =mn). The variance of the mean of the results obtained from the m composite samples would be:

$$\sigma_m^2 = (\sigma_0^2\delta + \sigma_a^2)/m + \sigma_0^2(1-\delta)/k = \sigma_0^2(1+\delta(n-1))/k + \sigma_a^2/m \qquad (8)$$

where, δ now refers to the intra-composite correlation, σ_0^2 is the between sample variance, ignoring measurement variance, and σ_a^2 is the pure analytical measurement (referred to as repeatability) variance. From Equation 8, it is seen that it is desirable that δ be small, which would be the case if it could be expected that true differences of levels between composite samples be negligible. Stratifying the population being sampled or selecting systematically from every m^{th} sample to form composite samples (for example, from 12 samples, selecting the

first, fourth, seventh and 10th as the first composite, and so forth, for 3 composite samples consisting of 4 samples each) would effectively minimize the value of δ. Assuming δ is small and can be ignored in Equation 8, the variance of the mean would depend upon the relative magnitude of σ_0^2 and σ_a^2; if m (the number of composites) is small, then, even with n being large, the variance of the mean could be large since the term σ_a^2 /m could be large.

However, often microbiological analyses are not able to handle large samples, and thus there may be a limit to the size of the composite samples. The limiting factors regarding the size of composite samples are the container size required for a (for example) 1/10 dilution, the ability to homogenize large samples and incubator space. Some laboratories may be equipped to handle large size samples, using walk-in incubators and such; however, most laboratories do not have such equipment.

Consideration also needs to be given to the sensitivity of the analytical procedure as a function of the sample size. In other words, analyzing composite samples might introduce a bias if the sensitivity of recovery were affected by compositing. These considerations might lead to limiting the number of samples, n, within a composite sample. This in turn might make less innocuous the assumption of a small δ.

Suppose it is decided that M grams (or ml if liquid samples are being considered) is the size of the composite sample. That is, the number of individual samples, n, in a composite sample, times the weight, w, (or liquid volume) of each individual sample, nw, should be equal to M. The total number of samples, k = mM/w. Equation 8 for the standard error of the mean of m composite sample results becomes:

$$\sigma_m^2 = \frac{w\sigma_{0w}^2}{mM}(1+\delta(n-1)) + \frac{\sigma_a^2}{m} \qquad (9)$$

where the symbol σ_{0w} refers to the between-sample variance for samples of weight (or volume) w. As w decreases (and thus increasing the number of samples, k) it would be expected that σ_{0w} would increase. The relationship between the two quantities: w and σ_{0w} would need to be explored in order to design an optimal composite sampling plan.

While composite sampling can lead to decrease of costs of sampling, it should be pointed out that the results obtained from composite sampling can mask information concerning the distribution of the levels of the measurand within the population being sampled. Information of the distribution of levels might be important for evaluating process control and for risk assessments that are primarily concerned with estimating risks typically associated with (occasional) high levels of some pathogen in food. Hence, for designing sampling plans, an understanding of how the results might be used is needed.

Designing sampling plans thus requires knowledge of variance components related to measurement and sampling variability associated with the sampling unit. In the SPC document (Enclosure B), the discussion does not address the effects of measurement error explicitly;

rather the document concentrates on the issues related directly to SPC, and estimated variances would include the contribution due to measurement. The total variability (due to sampling and analytical measurement) should be known or estimated in order to rationally design sampling plans– regarding the number of samples, composites, and repeat analyses that might be needed - and for constructing realistic OC curves. Information concerning specifics of this analysis can be found by reading Enclosures A, as well as reviewing the references in each case.

References

Cohen Jr, A. C. 1959. Simplified estimators for the normal distribution when samples are singly censored or truncated. *Technometrics,* 1(3):217-237.

Juran, JM. 1951. Quality Control Handbook, Third edition, McGraw-Hill Book Company, NY.

Persson, T. and Rootzén, H., 1977. Simple and highly efficient estimators for a type I censored normal sample. *Biometrika,* 64:123-128.

Singh, A. and Nocerino, J., 2002. Robust estimation of mean and variance using environmental data sets with below detection limit of observations. *Chemometrics and Intelligent laboratory Systems*. 60:69-86.

United States Environmental Protection Agency (EPA). 2000. Assigning values of non-detected/non-quantified pesticide residues in human health exposure assessments. Office of the Pesticide Programs, item 6047.