CHAPTER 4

CONSIDERATIONS FOR INTERPRETATION OF DATA AND RESULTS

INTRODUCTION

An important issue with LCI results is whether two numbers are really different from one another. For example, if one product has a total system requirement of 100 energy units, is it really different from another product system that requires 110 energy units? If the error or variability in the data is sufficiently large, it cannot be concluded that the two numbers are actually different.

STATISTICAL CONSIDERATIONS

A statistical analysis that yields clear numerical answers would be ideal, but unfortunately LCI data are not amenable to this. The data are not (1) random samples from (2) large populations that result in (3) "normal curve" distributions. LCI data meet none of these requirements for statistical analysis. LCI data for a given sub-process (such as potato production, roundwood harvesting, or caustic soda manufacture, for example) are generally selected to be representative of a process or industry, and are typically calculated as an average of two to five data points. In statistical terminology, these are not random samples, but "judgment samples," selected so as to reduce the possible errors incurred by limited sampling or limited population sizes. Formal statistics cannot be applied to judgment samples; however, a hypothetical data framework can be constructed to help assess in a general sense the reliability of LCI results.

The first step in this assessment is reporting standard deviation values from LCI data, calculated by:

$$s = \sqrt{\frac{\sum (x_1 - x_{mean})^2}{n-1}},$$

where x_i is a measured value in the data set and x_{mean} is the average of n values. An analysis of sub-process data from Franklin Associates, Ltd. files shows that, for a typical sub-process with two to five different companies supplying information, the standard deviation of the sample is about 30 percent of the sample average.

In a typical LCI study, the total energy of a product system consists of the sum of many sub-processes. For the moment, consider an example of adding only two numbers. If both numbers are independent of each other and are an average of measurements which have a sample standard deviation, s, of 30, the standard deviation of the sum is obtained by adding the variances of each term to form the sum of the variances, then taking the square root. Variances are calculated by squaring the standard deviation, s^2, so the sum of the variances is $30^2 + 30^2 = 900 + 900 = 1800$. The new standard deviation of the sum is the square root of the sum of the variances, or $\sqrt{1800}$ = 42.4. In this example, suppose both average values are 100, with a sum of 200. If reported as a percent of the sum, the new standard deviation is 42.4/200 = 21.3% of the sum. Another way of obtaining this value is to use the formula $s_{\%} = \frac{s/x_{mean}}{\sqrt{n}}$, where the term $s_{\%}$ is defined as the standard deviation of n data points, expressed as a % of the average, where each entry has approximately the same standard deviation, s. For the example, then, $s_{\%} = \frac{30\%}{\sqrt{2}} = 21.3\%$.

Going back to a hypothetical LCI example, consider a common product system consisting of a sum of approximately 40 subsystems. First, a special hypothetical case is examined where *all of the values are approximately the same size, and all have a standard deviation of 30%*. The standard deviation in the result is $s_{\%} = \frac{30\%}{\sqrt{40}} = 4.7\%$. The act of summing reduces the standard deviation of the result with respect to the standard deviation of each entry because of the assumption that errors are randomly distributed, and by combining values there is some cancellation of total error because some data values in each component system are higher than the true values and some are lower.

The point of this analysis, however, is to compare two results, e.g., the energy totals for two different product systems, and decide if the difference between them is significant or not. To test a hypothetical data set it will be assumed that two product systems consist of a sum of 40 values, and that the standard deviation, $s_{\%}$, is 4.7% for each product system.

If there is statistical knowledge of the sample only, and not of the populations from which they were drawn, “t” statistics can be used to find if the two product totals are different or not. The expression selected is:

$\mu 1 - \mu 2 = x1 - x2 +/- t_{.025} s' \sqrt{\frac{1}{n1} + \frac{1}{n2}}$, where $\mu 1 - \mu 2$. is the difference in population means, x_1-x_2 is the difference in sample means, and s' is a pooled standard deviation of the two samples. For the hypothetical case, where it is assumed that the standard deviation of the two samples is the same, the pooled value is simply replaced with the standard deviation of the samples.

The goal is to find an expression that compares our sample means to “true,” or population, means. A new quantity is defined: $\Delta = (\mu 1 - \mu 2) - (x1 - x2)$, and the sample sizes are assumed to be the same (i.e., n_1=n_2). The result is $\Delta = t_{.025} s' \sqrt{\frac{2}{n}}$, where Δ is the minimum difference corresponding to a 95%

confidence level, s' is the standard deviation of the sum of n values, and $t_{.025}$ is a t statistic for 95% confidence levels. The values for t are a function of n and are found in tables. This expression can be converted to percent notation by dividing both sides by the average of the sample means, which results in $\Delta\% = t_{.025} s'\% \sqrt{\frac{2}{n}}$, where $\Delta\%$ is now the percent difference corresponding to a 95% confidence level, and s'% is the standard deviation expressed as a percent of the average of the sample means. This formula can be simplified for the example calculation by remembering that $s'\% = \frac{s\%}{\sqrt{n}}$, where s% is the standard deviation of each energy entry for a product system. Now the equation becomes $\Delta\% = t_{.025} s\% \frac{\sqrt{2}}{n}$. For the example, t = 2.0, s = 30%, and n = 40, so that $\Delta\%$ = 2.1%. This means that if the two product system energy totals differ by more than 2.1%, there is a 95% confidence level that the difference is significant. That is, if 100 independent studies were conducted (in which new data samples were drawn from the same population and the study was conducted in the identical manner), then 95 of these studies would find the energy values for the two product systems to differ by more than 2.1%.

The previous discussion applies only to a hypothetical and highly idealized framework to which statistical mathematics apply. LCI data differ from this in some important ways. One is that the 40 or so numbers that are added together for a final energy value of a product system are of widely varying size and have different variances. The importance of this is that large numbers contribute more to the total variance of the result. For example, if 20 energy units and 2,000 energy units are added, the sum is 2,020 energy units. If the standard deviation of the smaller value is 30% (or 6 units), the variance is 6^2 = 36. If the standard deviation of the larger number is 10% (or 200), the variance is 200^2 = 40,000. The total variance of the sum is 36 + 40,000 = 40,036, leading to a standard deviation in the sum of $\frac{\sqrt{(40036)}}{2020}$ = 9.9%. Clearly, the variance in the result is much more greatly influenced by larger numbers. In a set of LCI energy data, standard deviations may range from 10% to 60%. If a large number has a large percentage standard deviation, then the sum will also be more uncertain. If the variance of the large number is small, the answer will be more certain. To offset the potential problem of a large variance, Franklin Associates goes to great lengths to increase the reliability of the larger numbers, but there may simply be inherent variability in some numbers which is beyond the researchers' control.

If only a few numbers contribute most of the total energy in a system, the value of $\Delta\%$ goes up. This can be illustrated by going back to the formula for $\Delta\%$ and calculating examples for n = 5 and 10. From statistical tables, the values for $t_{.025}$ are 2.78 for n = 5, and 2.26 for n = 10. Referring back to the hypothetical two-product data set with s% = 30% *for each entry*, the corresponding values for $\Delta\%$ are 24% for n = 5 and 9.6% for n = 10. Thus, if only 5 numbers out of 40 contribute most of the energy, the percent *difference* in the two product system energy values must increase to 24% to achieve the 95% confidence level that the two values are different. The minimum difference decreases to 9.6% if there are 10 major contributors out of the 40 energy numbers in a product system.

CONCLUSIONS

The discussion above highlights the importance of sample size, and of the variability of the sample. However, once again it must be emphasized that the statistical analysis does not apply to LCI data. It only serves to illustrate the important issues. Valid standard deviations cannot be calculated because of the failure of the data to meet the required statistical formula assumptions. Nevertheless, it is important to achieve a maximum sample size with minimum variability in the data. Franklin Associates examines the data, identifies the large values contributing to a sum, then conducts more intensive analysis of those values. This has the effect of increasing the number of data points, and therefore decreasing the "standard deviation." Even though a calculated standard deviation of 30% may be typical for Franklin Associates' LCI data, the actual confidence level is much higher for the large values that control the variability of the data than for the small values. However, none of this can be quantified to the satisfaction of a statistician who draws conclusions based upon random sampling. In the case of LCI data, it comes down to a matter of professional judgment and experience. The increase in confidence level resulting from judgment and experience is not measurable.

It is the professional judgment of Franklin Associates, based upon over 25 years of experience in analyzing LCI data, that a 10% rule is a reasonable value for $\Delta\%$ for stating results of product system energy totals. That is, if the energy of one system is 10% different from another, it can be concluded that the difference is significant. It is clear that this convention is a matter of judgment. This is not claimed to be a highly accurate statement; however, the statistical arguments with hypothetical, but similar, data lend plausibility to this convention.

We also conclude that the weight of postconsumer solid waste data can be analyzed in a similar way. These data are at least as accurate as the energy data, perhaps with even less uncertainty in the results. Therefore, the 10% rule applies to postconsumer solid waste weight. However, we apply a 25% rule to the solid waste volume data because of greater potential variability in the volume conversion factors.

Air and water pollution and industrial solid waste data are not included in the 10% rule. Their variability is much higher. Data reported by similar plants may differ by a factor of two, or even a factor of ten or higher in some cases. Standard deviations may be as high as 150%, although 75% is typical. This translates to a hypothetical standard deviation in a final result of 12%, or a difference of at least 25% being required for a 95% confidence of two totals being different if 10 subsystems are major contributors to the final results. However, this rule applies only to single emission categories, and cannot be extended to general statements about environmental emissions resulting from a single product system. The interpretation of environmental emission data is further complicated by the fact that not all plants report the same emission categories, and that there is not an accepted method of evaluating the relative importance of various emissions.

It is the intent of this appendix to convey an explanation of Franklin Associates' 10% and 25% rules and establish their plausibility. **Franklin Associates' policy is to consider product system totals for energy and weight of postconsumer solid waste weight to be different if there is at least a 10% difference in the totals. Otherwise, the difference is considered to be insignificant. In the detailed tables of this report there are many specific pollutant categories that are variable between systems. For the air and waterborne emissions, industrial solid waste, and postconsumer solid waste volume, the 25% rule should be applied.** The formula used to calculate the difference between two systems is:

$$\%\ \text{Diff} = \left(\frac{x-y}{\frac{x+y}{2}} \right) \text{X } 100,$$

where x and y are the summed totals of energy or waste for two product systems. The denominator of this expression is the average of the two values.