Draft
February 1, 2002

Statistical Analysis of Mercury Emission Rate Database

[Prepared by John Holmes for Utility MACT ranking subgroup]

Summary

In an attempt to array the EPA data for purposes of first identifying the best performers and second identifying a potential MACT floor, the subgroup identified a need to assess the variability that is apparent on the face of the data. The coal fired powerplant mercury (Hg) emission test database consists of 80 sets of test results consisting of three tests on each unit. The subgroup determined that the average 95% confidence interval in estimating the average emission rate of a unit in the EPA data set is plus or minus 30%. In addition to this variability factor, because it appears that the testing was not random, it's unclear whether the data accurately represent the variability in emissions that is present over the full range of operating conditions. As next steps, the subgroup believes that the full workgroup should work with EPA to consider whether the subgroup's analysis is accurate, learn how EPA has handled such data variability in other MACT proceedings, and determine how variability should be factored into this standard setting process.

This paper applies statistical analytical techniques to the ICR database on coal-fired utility mercury (Hg) emissions. The paper begins with a discussion of the statistical calculations that can be used to estimate the mean Hg emission rate of an individual unit and the confidence interval associated with the mean value. Using these calculations, it then looks more generally at the value of the database for determining the actual Hg emissions of the sampled units. Finally, it presents two important caveats.

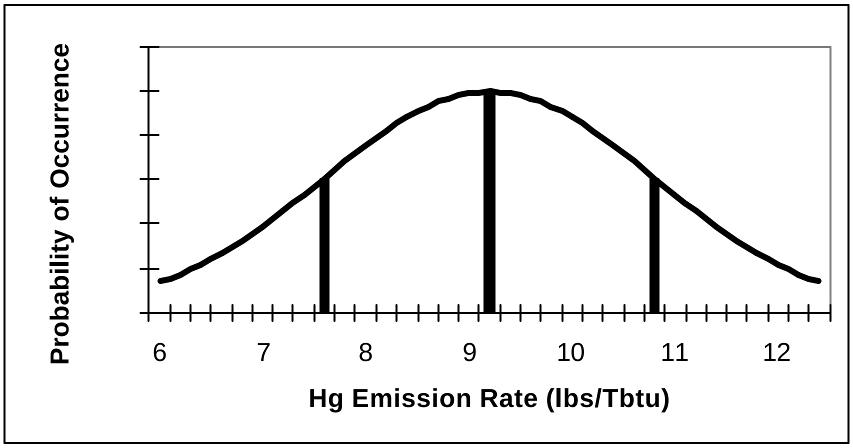
1. Determining Performance for an Individual Unit

Let us begin by reviewing how sampling methods are used to infer the (unknown) characteristics of a population. Gibson Generating Station Unit 3 was tested three times during 10/1999 with reported Hg emission rates of 8.1, 9.8, and 11.4 lbs/Tbtu (F factor adjusted). In this case, the *population* is the unit's Hg emission rate under all possible test times and conditions. The *sample* consists of these three tests, involving specific times and conditions.

Under certain assumptions, specifically, that the test results measure what they purport to measure and that the times/conditions of the tests are chosen randomly from all that are possible, we can use the statistics *of the sample* to estimate the characteristics of the *population* – i.e., the actual Hg emission rate of the unit in question. If we believe the population follows a normal distribution, we estimate its characteristics from sample values as follows:

- The population mean value is estimated as the mean of the sample, which is calculated as: X_o = SUM(X_i) / N, where i denotes a test and N is the number of tests

Figure 1: Distribution of Gibson Test Values During 10/1999



- The population variance is estimated by the calculation: Variance = SUM($(X_i - X_o)^2$) / (N - 1)

- The population standard deviation is estimated as: Sigma = SQRT(Variance)

For the Gibson 10/1999 tests, we estimate $X_o = 9.75$, Sigma = 0.94, and Variance = 0.89. Based on these statistics, the *inferred population* of test outcomes for Gibson is shown as the smooth bell-shaped curve in Figure 1; the bars show the values obtained in the three tests that were made. Clearly, a wide range of test outcomes appears to be possible, from a low value of about 6 lbs/Tbtu on the left to twice that amount on the right. Note that several factors can contribute to the range of variation seen in the graph: changes in the Hg content of the coal between tests, variation in the Hg removal efficiency of the controls, differences (if any) in unit operating conditions between the tests and sampling and test method variability. Note also that the range of variation seen in these three data points could understate the actual range of variation in the real world, if the approach to these three tests tended to exclude times or operating conditions under which higher or lower emission rates would be measured.

What matters most for MACT standard setting is unit performance averaged across the time period (and associated operating conditions) on which the standard is to be set. This time period has not been established yet, so let us focus attention on the overall mean performance. If we were to perform the 3-test experiment a number of times on a unit, we would get different estimates for the population mean value each time. The uncertainty (standard deviation) in the mean value, termed its standard error, is computed as follows:

- Var_{mean} = Variance / N

- $\text{Sigma}_{\text{mean}}$ = SQRT (Var_{mean}) = Sigma / SQRT(N)

If the data are normally distributed, then the estimated mean value follows the T distribution with N-1 degrees of freedom. For large sample sizes (N > 30), the T distribution approaches the normal distribution, and the 95 percent confidence interval for a mean value is approximately +/- 2 * $\text{Sigma}_{\text{mean}}$ (a 2-sigma error band). If the sample size is small, we must determine the multiplier from a table of T values, and we must take care to assure that the data are truly normally distributed. For N=3, as is the case in this database, the multiplier for a 95 percent confidence interval is 4.3, or more than twice that of a case where the sample is large.

Let us compute the confidence interval for the Gibson 10/1999 tests. $\text{Sigma}_{\text{mean}}$ equals 9.75 and the 95 percent confidence interval for the unit is:

$$95 \text{ percent CI} = 9.75 \ +/- \ 4.3 * 0.94 = (5.7,\ 13.8) \text{ lbs/Tbtu}$$

This means that we have a 95 percent confidence that the actual mean is between 5.7 lbs/Tbtu and 13.8 lbs/Tbtu. The confidence interval can also be expressed, often more conveniently, as a percentage of the mean value. In the above example, one would say that the mean value of 9.75 lbs/Tbtu is known to within +/- 4.3*0.94/9.75 = +/- 41 percent.

The uncertainty in this unit's measured average performance is so large that we can say, with 95 percent confidence, only that its mean Hg emission rate lies somewhere in the range shown in Figure 1. That is, *the uncertainty in the mean value is as large as the variation in the individual test results.* In fact, this will necessarily be true for any unit subjected to only 3 tests[1]. Further, Gibson is not the most extreme case in the database: 8 other units in the database have greater variation in their test results and wider confidence intervals for their estimated mean emission rates. Three tests are simply too few to determine a unit's true average performance with good confidence.

2. <u>Performance Measures for the Full Sample of Units in the Database</u>

Let us generalize what was seen above by applying the same analysis to the entire sample of generating units in the ICR database. The mean Hg emission rate and its 95 percent confidence interval have been computed for each of the 80 units in the sample. Figure 2 displays these values as a scatter plot with the mean emission rate on the horizontal axis and the 95 percent confidence interval, expressed as a percent of the mean value, on the vertical axis. Each symbol is a single generating unit. Because the units with baghouse particulate controls are considered among the best controlled and are likely to be most critical to the standard setting process for units in their category or subcategory, they are highlighted in the figure using the enlarged, open symbols.

Plotted this way, a unit's estimated mean emission rate determines its horizontal position on the graph, while the percentage uncertainty in the emission rate determines its vertical position. Points far above

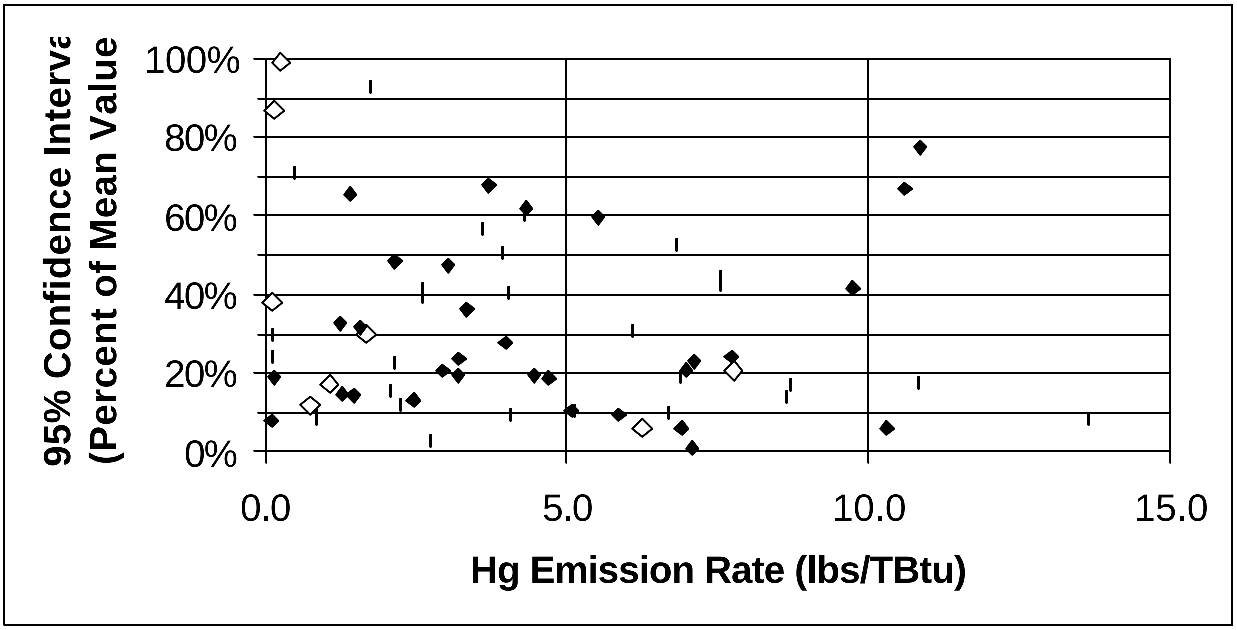
[1] See the appendix at the end of this paper for a demonstration of this result.

the horizontal axis have large uncertainties in the emission rate, while points close to the horizontal axis have small uncertainties. If the quality of the database were sufficient to determine the average Hg emission rate for each unit to within +/- 10 percent of the mean (95 percent confidence interval), then *all* of the points would lie below the horizontal gridline located at 10% on the vertical axis.

One sees the following in the chart:

- The 95 percent confidence intervals are greater than +/-10 percent for the large majority of units. Only two are within +/- 5 percent, and only 14 units are within +/- 10 percent.

Figure 2: Mean Hg Emission Rate and 95% Confidence Intervals



- The median confidence interval is about 30 percent. That is, for half of the units, the 95 percent confidence interval for the mean emissions rate is no better than +/- 30 percent.

- The confidence interval, expressed in percentage terms, can be wide (i.e., poor) for units with small average emission rates

Of the 15 units with mean emission rates below 1.0 lbs/Tbtu, only two units have a 95 percent confidence interval as small as 10 percent of the mean. Taken as a group, the 15 units have an average emissions rate of 0.25 lbs/Tbtu with a 95% confidence interval of +/- 0.17 lbs/Tbtu (68 percent of the mean). Thus, we can say with 95 percent confidence that the average for this group lies between emission rates of 0.08 and 0.42 lbs/Tbtu.

The generating units with baghouse particulate controls (open symbols) tend to mirror the trends seen in the data overall. At any average emission rate, the confidence intervals for baghouse units are similar to those for other units. Analysis of the baghouse subgroup will have an uncertainty that is similar to the database overall, or perhaps worse when compounded by the smaller size of the baghouse sample.

The wide range in the performance values and uncertainties seen in Figure 2 indicate that the 3-test database is unlikely to be adequate for supporting standards development with an acceptable level of statistical confidence. The problem is that 3 tests are too few, and the only completely satisfactory solution to the problem is to obtain more tests. Additionally, as discussed at the end of this paper, the confidence intervals shown above may be *artificially small* due to the potential for biased sampling in the generation of this data.

On the other hand, it *may* be possible to use the existing data to identify subsets of apparently high-performing units, on which further emissions testing can be performed. The accuracy of this determination (that is, whether a particular unit really performs better than the one below it in a ranked list) will depend on the sizes of $\text{Sigma}_{\text{mean}}$ for the individual units compared to the size of the difference between the two consecutive units. There are statistical procedures for testing the difference between two mean values, although this is difficult to do properly in small samples.

A useful (although rough) rule-of-thumb is that two means can be said to truly differ from each other when the +/- one-sigma confidence intervals of the means do not overlap. That is, if X_o minus $\text{Sigma}_{\text{mean}}$ for the larger observation is still greater than X_o plus $\text{Sigma}_{\text{mean}}$ for the smaller observation, we can say (roughly) that the two observations are truly different. A cursory examination of unit ranking using this one sigma rule-of-thumb suggests that it may be possible to rank the units in order of increasing average emission rate with at least a reasonable level of statistical confidence that the ranking is correct.

Thus, one strategy may be to identify the high-performing units in a particular category or subcategory using the ICR database, then conduct additional, more extensive testing on those units to determine their actual emissions over a range of conditions, and use that data to derive the "floor" for the standards.

3. Important Caveats to the Use of the EPA Data

The foregoing analysis was premised on two very important assumptions regarding the existing data, specifically that:

- The test values measure what they purport to measure (actual Hg emission rates), and

- The operating conditions of the three tests performed on each unit were chosen randomly from all that are possible.

If both assumptions are true, then we may take the data as a fair and unbiased sample of the population and properly use sample statistics to infer the characteristics of the population. If one or both assumptions are false, then the statistical inferences are not strictly valid, the data are potentially misleading (at best), and conclusions drawn from the data are potentially wrong. Two caveats warrant careful consideration and assessment *before* the existing data are used in statistical analysis.

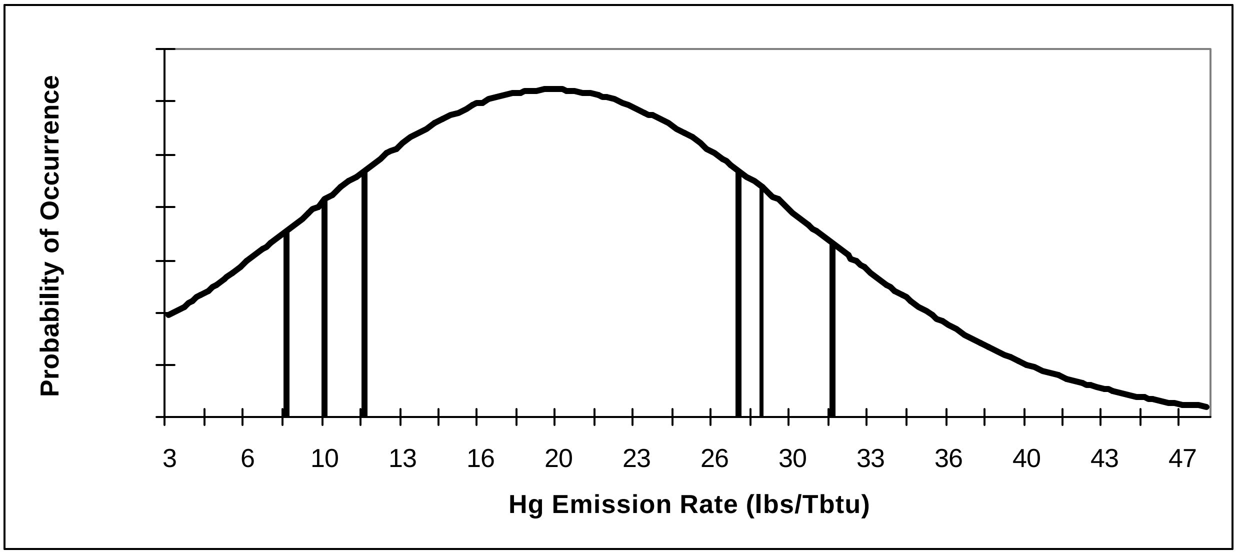
1. A prior review of the data showed that some 22 percent of the tests indicate an *increase* in Hg emissions across the last control device. It *must* be resolved whether such results have physical meaning or not and what their presence says about the quality of the data for those units and the database in general. If the result is a legitimate outcome, then we must understand how such data can be used in assessing unit performance. If the result is not legitimate, then erroneous or inconsistent data must be corrected or deleted from the data set, and all other data values should be scrutinized for problems akin to those that caused the erroneous results.

2. An assessment *must* be made of the potential for bias in the data as a result of EPA's testing protocol, which did not specify the unit operating conditions under which tests were performed and which left each facility free to choose when the test took place. If the data are not a fair, balanced sample of all of the times and conditions under which units operate, then the data cannot accurately measure unit performance in the real world. If the data are affected by sampling bias of this kind, then statistical measures such as confidence intervals are very likely to *understate* the actual uncertainty present in the data.

The latter caveat is not merely a theoretical concern. This can be seen in the one instance where a unit was subjected to a repeat sequence of three tests. During 03/2000, testing was repeated at Gibson Generating Station Unit 3. All of the new tests gave emission rates far in excess of the original test values[2]. When added to the earlier testing, this produces the distribution of test values shown in Figure 3. Compare the width of the distribution to that of Figure 1. The standard deviation was 1.6 lbs/Tbtu in the original testing, but increases to 10.7 lbs/Tbtu when all tests are pooled. The second set of tests has increased the overall variability in the sample by including data recorded under operating conditions not encountered in the first set of tests.

[2] The unit operators believe the high emission rates were caused by soot blowing during the three retests in 3/2000. Hg emissions, both before and after the final ESP-CS control device, exceeded the Hg content of the coal. The operators believe that Hg, which had been deposited with soot formed in prior operation, was released as a result of soot blowing and recorded by the tests.

Figure 3: Distribution of All Testing for Gibson Unit 3



This is but one example, related to a specific operating condition, that illustrates the general rule: *if a sample does not fairly reflect all operating conditions that may occur, the sample data will tend to understate the variability actually present in the real world*. This demonstrates the importance of determining whether the data show evidence of bias in terms of the times and conditions under which the tests were conducted. Assume, for example, that Hg emissions are low under high-load operation, but increase significantly under part-load conditions or load swings. Then, if the tests tended to be performed under high-load stable conditions, the existing data can substantially under-estimate Hg emissions from the units and give little or no insight into control system performance under actual operating conditions.

At least limited data are available in the test reports filed with EPA by unit operators on the operating conditions under which the tests were conducted, including date and time of day (indirect measures of operating conditions), unit load (MW net), steam flow, furnace exit gas temperature, and CEM data. A first step in a bias assessment would be to examine the distributions of the tests in terms of day of week, time of day, and unit load factor. Unless the data are well dispersed across the range of time and operating condition values likely to be encountered, one should assume that the potential for bias is high.

Appendix:

Confidence Intervals for Units with Three Tests

Let Sigma equal the standard deviation of the population of test results. Then, +/- 2*Sigma gives us, approximately, the range within which 95 percent of the population will lie.

The standard error of the mean value, as estimated from N tests drawn from the population, will be Sigma / SQRT(N). For N = 3, the 95 percent confidence interval of the mean is:

95% CI = +/- 4.3 * Sigma / SQRT(3)

= +/- 4.3 * Sigma / 1.73

= +/- 2.5 * Sigma

We see, therefore, that the confidence interval of the mean is actually somewhat wider than the range that encompasses 95 percent of the data. Three tests are the *bare* minimum needed to estimate the population mean and standard deviation, but three tests are too few to do so with any real degree of statistical confidence.