ADDRESSING DATA QUALITY ISSUES THROUGHOUT THE SITE CHARACTERIZATION PROCESS TO MINIMIZE DECISION ERRORS

D.W. BERMAN, Aeolus, Inc, Albany, California*

ABSTRACT

Making an incorrect decision concerning the need for cleanup at a hazardous waste site can result either in leaving a potentially harmful situation in place or unnecessarily committing large sums of money to cleaning up a site that otherwise poses no significant threat. Site management decisions are unavoidably associated with some degree of error because: (1) decisions are always rendered based on incomplete information and (2) data used to support decisions are subject to uncertainty. However, there are also many questionable practices that are commonly employed throughout the site characterization process that increase decision error beyond the unavoidable minimum.

A variety of common practices that can contribute substantially to decision error are identified below and evaluated both to identify conditions under which they lead to erroneous conclusions and to quantify the rate that such errors can occur. It will also be shown how these problems can be avoided by paying close attention to data quality issues throughout the site characterization process including, particularly, use of the data quality objectives (DQO) process to focus the planning of site investigations. Importantly, the problems introduced by the practices discussed are insidious; unless one addresses data quality issues throughout the site characterization process, one will never know whether a decision error has been committed.

INTRODUCTION

Under the Federal Superfund program, state analogs to the superfund program, and (increasingly) hazardous waste sites in general, site risk assessments are employed to determine whether hazardous waste sites pose unacceptable risks to human health and/or the environment and, therefore, whether cleanup is warranted. Cleanup costs commonly reach tens of millions of dollars at the largest of such sites. Thus, making an incorrect decision can result either in leaving a potentially harmful situation in place or unnecessarily committing large sums of money and limited resources to cleanup a site that otherwise poses no significant threat. It is therefore important to minimize errors in the procedures used to determine the need for cleanup.

---

* D. Wayne Berman, Aeolus, Inc., 751 Taft St., Albany, CA 94706; e-mail: bermanw@comcast.net.

The conclusions of a risk assessment are typically drawn by comparing estimated risks either to each other or to fixed targets. Comparisons between estimated risks, for example, may be used to rank the magnitude of problems at multiple sites. When fixed targets are selected to represent maximum acceptable risk, comparisons between estimated risks and fixed targets are typically performed to determine whether risks posed by a particular site (or source) are acceptable. When risks are found to be unacceptable, cleanup is generally required. For the indicated comparisons, the probability that conclusions from a risk assessment are "correct" is a function of the degree to which risks estimated in the assessment are accurate and precise or, at a minimum, the degree to which any biases introduced in the risk assessment process are consistent across risk estimates and target values.

In turn, the degree to which estimated risks are accurate and precise (or the extent to which biases are consistent) are functions of the quality of the underlying data and the appropriateness of the procedures used to derive risk estimates from such data. Over the years, however, several practices have been adopted that may compromise the quality of the data used to support risk assessments so that the conclusions drawn from such data should be considered unreliable. Although a numerical result can always be calculated by a manipulation of data, there is no assurance that such a result bears any relationship to true conditions in the field, unless one considers data quality. The problem is insidious.

When evaluating data to support risk assessments and render site management decisions, some degree of error is unavoidable. This is because:

(1) decisions are always rendered based on incomplete information; and

(2) data used to support decisions are subject to uncertainty.

However, there are also many questionable practices that are commonly employed throughout the site characterization process that can increase decision error beyond the unavoidable minimum. These include use of procedures that generate data that are not optimal for the purposes to which the data are applied and use of statistical procedures to evaluate data that are not appropriate for the characteristics of the data being evaluated. Questionable practices that result in the use of non-optimal data include:

- using data generated in a study designed for one purpose in an evaluation intended for a different purpose;
- combining data from multiple investigations without considering the compatibility of the multiple data sets for a given purpose;
- designing sampling plans that generate unintentionally biased data;

- using data from biased investigations to develop estimates that are intended to be unbiased; and

- using inappropriate procedures to "adjust" data sets to account for non-detects.

Questionable practices that result in the application of statistical procedures that are inappropriate for the characteristics of the data being evaluated include:

- not properly matching decision rules to the characteristics of the data before proceeding with an evaluation;

- using "default" procedures without due consideration of their appropriateness to a particular situation; and

- misapplying the laws of inequalities when constructing decision logic.

In recent years, several colleagues and I have evaluated the kinds of problems that can be created by carelessly employing the practices listed above. In a series of recent publications, we have characterized the conditions under which such practices lead to erroneous conclusions and, in some cases, to quantify the rate that such errors can occur. An overview of the results of these studies and their implications for site characterization is presented in the Discussion Section below. This follows a brief background discussion to introduce relevant terms and concepts.

BACKGROUND

Before exploring the ways in which questionable practices can contribute to decision errors, it is necessary to introduce the formalism typically employed when using field measurements to support field data. The types of decision errors that can occur and the manner in which they are controlled are also introduced.

Decision Rule Formalism

In most site risk assessments, the decision whether to remediate ultimately reduces to determining whether the arithmetic mean, "m," of some contaminant concentration within a defined area or volume element of the environment exceeds a critical target value, "Q" (USEPA 1992). In this context, m is the *true* arithmetic mean of the contaminant concentration and Q typically represents the boundary between acceptable and unacceptable conditions (i.e., the boundary between acceptable concentrations and those that are large enough to present unacceptable risks). The manner in which such comparisons fit into the larger scheme of a risk assessment has been described previously (Berman 1995).

The comparison between Q and estimates of m represents a test in which a *null* hypothesis (e.g. that $m < Q$, which might indicate that a site is "clean") is assumed to be true unless conditions at the site warrant rejecting the null in favor of an alternate hypothesis (e.g. that $m \geq Q$, which might indicate that cleanup is required). It is equally valid to begin with the assumption that the site is "contaminated" (i.e. that $m > Q$) and to test whether this null can be rejected in favor of an alternate (i.e. that $m \leq Q$) to "prove" that a site is clean. In fact, current guidance favors this latter formulation when making decisions at Superfund sites (USEPA 1992). As shown in the Discussion Section, however, these two hypothesis formulations are equivalent so that the probability of reaching an incorrect conclusion is not affected by the choice of the formulation of the null hypothesis.

In the absence of error (uncertainty) in an estimate of m, a direct comparison between m and Q provides a perfect test of the null hypothesis. However, the true value of m can never be known exactly. This is due primarily to the limited ability of a finite number of samples that are collected from a large and variable environmental medium to completely represent the characteristics of that medium. Also, methods employed for sampling and analysis are subject to error and this introduces additional uncertainty in the estimates of m. Thus, summary statistics that are biased estimates of m (such as upper or lower confidence limits) are typically derived from field measurements and compared to Q to determine *with some predefined level of confidence* whether m is likely to exceed Q. In fact, a variety of summary statistics can be (and have been) used to test hypotheses concerning the true mean of a field distribution of concentrations and each offers various advantages and limitations (Berman et al. 1998a and 1998b).

The nature of the summary statistic selected to represent the mean, m, the manner in which it is calculated from field measurements, and the formulation of the hypotheses employed to compare that statistic to the target, Q, constitute a "decision rule." The probability that a decision rule leads to an incorrect decision (i.e. the error rate) is a function of the characteristics of the decision rule, the nature of the underlying true condition at the site and the degree to which the set of measurements employed to evaluate the decision are representative of the environmental medium that is the subject of the decision.

The degree to which a set of samples can be assumed to represent the sampled medium is a strong function of the process by which locations were selected for the samples collected, the number of samples collected, and the characteristics of the methods employed for sampling and analysis. The effects on decision errors associated with the data quality limitations imposed by each of these factors were explored in a previous study (Berman 1995) and are summarized in the Discussion Section under considerations associated with non-optimal data sets.

Even assuming that data are collected optimally, however, the choice of the decision rule can also contribute to decision error. The manner in which the choice of decision rule affects decision error has also been explored in previous studies (Berman et al. 1998a and 1998b) and is also summarized in the Discussion Section under considerations associated with use of inappropriate statistical procedures.

Types of Decision Errors

Given a particular decision rule in which, for example, it is assumed that $m < Q$ as a null hypothesis and a comparison between Q and some summary statistic employed to represent m is used to decide whether to accept or reject the null hypothesis, there are four possible outcomes:

- m is truly less than Q and the comparison leads properly to a conclusion that m is less than Q;
- m is truly less than Q but the comparison leads falsely to a conclusion that m exceeds Q;
- m truly exceeds Q and the comparison leads properly to a conclusion that m exceeds Q; or
- m truly exceeds Q but the comparison leads falsely to a conclusion that m does not exceed Q.

The first and third of the above listed outcomes are correct results. The second and fourth are erroneous. Thus, there are two kinds of errors that can potentially occur during the application of a decision rule. A "Type I" error occurs when the null hypothesis is falsely rejected. In the illustration presented above, this corresponds to the second of the listed outcomes. A "Type II" error occurs when the null hypothesis is falsely accepted and corresponds to the last of the possible outcomes listed in the above illustration.

Note that, because Type I and Type II errors are defined based on whether the null hypothesis is accepted or rejected, the relationship between these types of errors and the specific outcomes of a comparison depends on the way that the null hypothesis is defined. Thus, for example, if a new decision rule is examined in which the null hypothesis is assumed to be $m > Q$ (the reverse of the illustration above), then the second of the four outcomes listed above becomes a Type II error and the fourth becomes a Type I error.

In the face of uncertainty, formal decision rules are applied to control decision error. Statistical control is achieved when, due to the inherent design of a particular statistical test, the null hypothesis is rejected at a *fixed and defined* rate at the decision point (i.e. when $m = Q$). The rate that the null is rejected at the decision point is termed the significance or α (alpha) level for the test.

Importantly, statistical control implies that the significance level of the test is independent of sample size and other characteristics of the data to which the decision rule is applied. However, the rate of rejection of the null hypothesis for true conditions other than the decision point (i.e. when $m \neq Q$), is generally a function of both the true condition (i.e. the value of m) and the sample size. The rate at which the null hypothesis is rejected for conditions under which rejection is correct (i.e. when $m \geq Q$) is termed the power of the test. If a statistical test is well behaved, the power of the test increases as sample size increases.

In practice, the intended statistical control of a particular test is achieved when the characteristics of the data to which the test is applied completely satisfy the validity criteria for the test. The validity criteria for a particular statistical test are the mathematical axioms that were employed to construct the test (e.g. some tests require that data distributions be symmetric). However, tests are frequently applied to data that exhibit characteristics that do not strictly satisfy such criteria. If deviations are small, the test may still behave reasonably and statistical control will be maintained. However, the characteristics of environmental data frequently fall at the fringe of acceptability for the kinds of spatially-independent statistical tests that are commonly applied. The conditions under which statistical control fails for various commonly employed statistical tests and estimates of the resulting error rates have been reported in the literature (Berman et al. 1998a and 1998b) and are summarized in the Discussion Section below.

DISCUSSION

The manner in which each of the questionable practices identified in the Introduction can contribute to decision errors is considered below.

The Consequences of Using Non-optimal Data Sets to Support Decisions

The first five of the questionable practices listed in the Introduction lead to use of non-optimal data to support decisions and the consequences of using non-optimal data are illustrated in a simulation that was previously reported in the literature (Berman 1995). The advantage of studying a simulated case (versus an actual case) is that the true conditions for the simulated case can be defined absolutely. Thus, the reliability of the conclusions drawn from characterization of a simulated site can be determined by comparing such conclusions to the underlying true conditions. For real sites, in contrast, the underlying true conditions can never be known with certainty.

In the simulation reported by Berman (1995), a hypothetical site was constructed in which an old landfill is rumored to exist at the end of an old, paved road. It is further assumed that trucks from a nearby plant would drive to the end of the paved road and dump a hazardous "Compound Q" at the edge of the landfill. A map of the hypothetical site is presented in Figure 1.

The underlying "true condition" was constructed for this simulation by defining a two-dimensional distribution of contamination that is consistent with the above-stated scenario. This distribution was then "sampled" by picking sampling locations and setting the value of the resulting "measured" concentration to the value of the concentration that exists within the synthesized distribution at the location sampled.

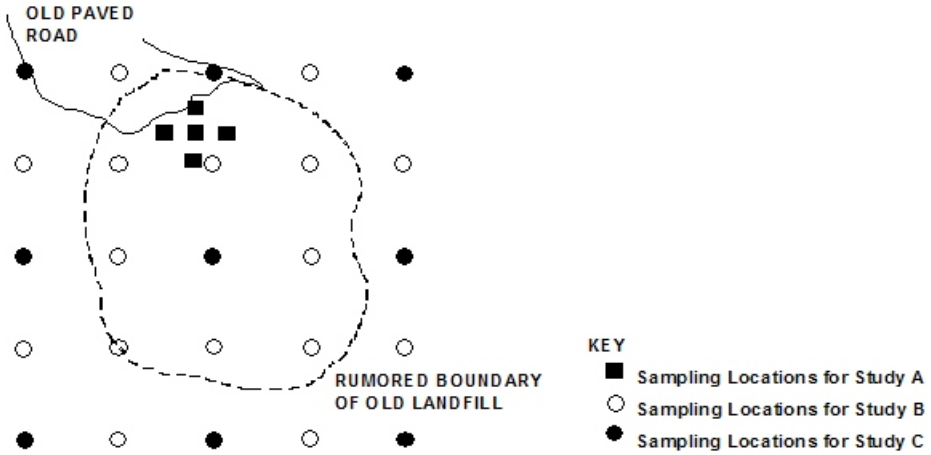

FIGURE 1 Sampling Locations for a Hypothetical Landfill (Source: Berman 1995)

In a sequence that parallels the evolutionary path common to investigations of real sites, it is assumed in this simulation that the site is ultimately sampled three separate times for three separate purposes:

(1) to confirm the presence of the rumored Compound Q, a series of samples are collected from locations that are selected purposively in areas expected to be most highly contaminated (Study A). This creates a data set that is intentionally biased high;

(2) to estimate the "nature and extent" of contamination, the area within which the landfill is assumed to reside is sampled systematically (Study B); and

(3) to improve characterization of the "nature and extent" of contamination (to support remedial design), sampling is conducted over a systematic grid that is more closely spaced than that sampled in Study B (Study C).

The locations of the samples collected during each of the studies listed above are also depicted in Figure 1. Estimated concentrations derived using the measurements from each of the three studies are summarized in Table 1.

As indicated in the first column of Table 1, the data sets evaluated include: results from Study A, results from Study B, a data set consisting of the pooled results from the combined Studies A and B, and results from Study C. The number of samples in each data set is listed in the second column of the table. The third column of Table 1 indicates the results of goodness-of-fit tests conducted on each data set to determine whether it might be adequately described by a normal or a lognormal distribution; all of the data sets listed can be adequately described by lognormal distributions.

TABLE 1 Summary of Findings for the Data Sets from Studies of a Hypothetical Landfill (in ppb)[a]

Data Set	No. of Samples	Statistical Distribution[b]	Mean (generic)[c]	MLE Mean (log-dist)[d]	UCL_{95} (log-dist)[e]	CV[f]
Study A	5	Lognormal	$4.7x10^5$	$6.8x10^6$	$1.8x10^8$	2.2
Study B	9	Lognormal	$3.2x10^3$	$2.1x10^5$	$2.9x10^6$	1.98
Pooled A and B	14	Lognormal	$1.7x10^5$	$1.3x10^7$	$1.6x10^8$	3.64
Study C	25	Lognormal	$9.0x10^4$	$9.3x10^5$	$6.1x10^6$	3.4

[a] Adapted from Berman (1995).

[b] Based on Goodness-of-fit tests performed as described in Gilbert (1987).

[c] Arithmetic mean calculated using the sum of values divided by the number of values (Gilbert 1987).

[d] Arithmetic mean calculated using the simplified maximum likelihood estimator for a lognormal distribution (Gilbert 1987).

[e] Upper 95% confidence limit to the mean of a lognormal distribution (Land 1971).

[f] "CV" is the coefficient of variation, which is the appropriate independent variable for describing the variability of lognormally distributed data (Gilbert 1987).

The fourth and fifth columns of Table 1 provide "best" estimates of the mean of each data set calculated, respectively, using the general equation for the arithmetic average and the simplified maximum likelihood estimator (MLE) for the arithmetic mean of a lognormal distribution (Gilbert 1987). The sixth column of the table provides another kind of estimate for the mean for each data set. In this column, estimates are presented for the upper 95% confidence limit to the mean (UCL_{95}) of each data set. These were calculated using the Land Equation, which is appropriate for lognormally distributed data (Gilbert 1987). The last column of the table presents estimates of the coefficient of variation, which is a measure of the spread (variability) of the data (Gilbert 1987).

Several trends are immediately apparent from Table 1. It is clear, for example, that the estimates of the mean concentration for Compound Q vary by five orders of magnitude across the data sets for the kinds of estimates of the mean depicted (i.e., across all of the entries in Columns 4, 5, and 6). Even considering results based on any one procedure for estimating the mean (i.e., within each column), results vary by up to two orders of magnitude across data sets. Which is correct? The answer depends on what question is being addressed. For example, if the goal is to find the highest localized concentrations in the landfill, then mean estimates based on Study A may be appropriate.

If the goal is to determine the average concentration of Compound Q over the contaminated area of the landfill as a whole, Study C (the most dense of the data sets sampled systematically) provides an answer that is closest to the truth. The less dense of the data sets sampled systematically (Study B) also gives answers that are in relatively close agreement with Study C. In contrast, concentration estimates derived from the data set that was intentionally biased (Study A) can be as much as two orders of magnitude higher than that obtained from the unbiased Studies B or C. Concentrations estimated from the data set derived by pooling Study A and Study B are also as much as two orders of magnitude greater than the concentrations estimated for Study B alone.

It is also clear from Table 1 that estimates of the mean concentration for any specific data set vary by up to three orders of magnitude, depending on which procedure is employed for calculating the estimate (i.e., within any one row across Columns 4, 5, and 6). Because all of the data sets have been shown to be adequately described by lognormal distributions, the correct procedure for deriving a best-estimate of the mean is to use the MLE (Column 5). Although the general equation for estimating the mean (Column 4) is an unbiased estimator for any distribution, for lognormal distributions, it converges to the true mean more slowly than the MLE. Thus, when data are lognormally distributed, the traditional procedure for estimating the mean should not be used for small data sets (such as those evaluated in Table 1) because results will not be stable.

Estimates derived using the general equation for the arithmetic mean are presented in Table 1 to illustrate what can happen if the incorrect estimator for the mean is used during data evaluation. This commonly occurs, for example, if the characteristics of a data set are ignored when selecting the statistical procedure to be used for data evaluation. That the two best estimates of the mean derived for each data set (Columns 4 and 5 of Table 1) vary by an order of magnitude is a function of the lack of stability for the means estimated using the general equation (Column 4). The direction of the apparent bias between the two means is due simply to chance; for other data sets, the traditional mean can just as easily be orders of magnitude larger than the means calculated using the MLE equation (Gilbert 1987). Thus, biases introduced by the use of this type of incorrect statistical procedure will be inconsistent across risk estimates.

The UCL_{95} is also presented in Table 1 because such upper bound estimates of the mean are more frequently employed in risk assessments than best estimates of the mean. These estimates are used because they can provide formal statistical control of uncertainty (as defined in the Background Section). Use of the UCL_{95}, for example, assures that there will be no more than a 5% chance that the true mean is underestimated. Therefore, if such an estimate is compared to a target value in a risk assessment, there would be no more than a 5% chance that a site is falsely determined to be clean.

Importantly, the correct procedure for calculating a UCL_{95} varies depending on the characteristics of the data. Careless selection of an inappropriate procedure may result in calculating UCL_{95} estimates that are orders of magnitude different than the appropriate value. Clearly, therefore, such misuse can contribute substantially to decision error. The UCL_{95} estimates presented in Table 1 were derived using the appropriate procedure for lognormally distributed data, which appears appropriate for this case. The consequences of using inappropriate statistical procedures is discussed in a later subsection of this discussion.

The results in Table 1 indicate that estimates of concentrations derived from field measurements can vary by orders of magnitude depending on what data are used and how the data are evaluated. Clearly, employing inappropriately derived concentration estimates in a risk assessment can contribute substantially to decision errors. Given the above discussion, deriving appropriate concentration estimates requires that one employ data that are generated in investigations

designed specifically for the purpose for which the data are intended (i.e., requires use of optimal data). Thus, one should generally avoid:

- using data generated in a study designed for one purpose in an evaluation intended for a different purpose;
- combining data from multiple investigations without considering their compatibility for a given purpose;
- designing sampling plans that unintentionally generate biased data; and
- using data from biased investigations to generate estimates that are intended to be unbiased.

Thus, to minimize the chance of increasing decision errors due to the manner in which data are collected and evaluated, the first four of the questionable practices listed in the Introduction should generally be avoided. The fifth of the questionable practices, use of improper adjustments to account for non-detects, is discussed below.

Note, deriving appropriate concentration estimates also requires that the procedures used to evaluate such data be matched carefully both to the objectives of the evaluation and to the characteristics of the data. Contributions to decision errors associated with the selection of data evaluation procedures are discussed later in this Section.

Using of one-half of the Sample Quantitation Limits for Non-detects

It is common practice to substitute one-half of the sample quantitation limits for non-detects when evaluating a set of measurements to estimate concentration. The consequences of this practice were illustrated in a talk presented by Kenny Crump to the State of California Department of Toxic Substances Control (Crump 1993).

In his talk, Crump (1993) showed that, as long as more than a very small fraction (a few percent) of the measurements in a data set are non-detect, the mean of the data set will be overestimated whenever values of one-half of the sample quantitation limits are substituted for non-detects in the data set. This is particularly true when data sets exhibit behavior that is characteristic of lognormal distributions, which is relatively common for environmental data (Berman 1994).

Crump (1993) showed further that overestimation likely increases as the fraction of non-detects increases. In some of the examples presented in his talk, the bias was as much as several orders of magnitude. However, unless the mean is also calculated using a less biased procedure (such as the maximum likelihood estimation procedure discussed in the talk), the magnitude of the bias introduced in the mean by the use of this common substitution is unquantifiable. In addition,

because the bias varies as a function of the characteristics of the data, biases will be inconsistent across data sets and, correspondingly, across risk estimates.

The effect that use of one-half the sample quantitation limit has on upper bound estimates of the mean is even more confusing (Crump 1993). In this case, the bias introduced may be either positive or negative, depending on a number of factors and, once again, the magnitude is unquantifiable (unless an unbiased procedure for estimating the upper bound is also employed for comparison). Such biases are likewise inconsistent across risk estimates.

Given that use of one-half the sample quantitation limit can create errors in estimates of mean concentrations that can vary by orders of magnitude from estimates derived using more appropriate procedures, it is clear that indiscriminate use of these kinds of substitutions can contribute substantially to decision errors.

The Consequences of Using Inappropriate Decision Rules

Even assuming that the data to be used to support a risk assessment have been generated in an optimally designed investigation, the chance of committing a decision error can still be increased substantially depending on the nature of the decision rule employed during the evaluation of the data.

The vast majority of procedures employed for evaluating site data are based on what is called "parametric analysis." In a parametric analysis, data are assumed to display a distribution that closely mimics the characteristics of a known statistical distribution (e.g. a normal or lognormal distribution) with defined central value and spread. In turn, the characteristics of the known distribution are assumed to closely parallel the corresponding characteristics of the data.

When there is good correspondence between a data distribution and the shape of a known distribution, then the properties of the known distribution closely match the properties of the data. However, when the correspondence is poor, the properties of the known distribution are unlikely to relate well to the properties of the data. To minimize decision errors, it is therefore important to closely match the data to be evaluated to the characteristics of the known statistical distribution that is selected to represent the data. Sometimes this is done formally using goodness-of-fit tests and we recommend that such tests be employed routinely during the analysis of data from hazardous waste sites. Frequently, however, the relationship between a particular data set and the statistical distribution chosen to represent it is simply assumed and this can lead to substantial decision errors.

The quality of the relationship between a known statistical distribution and a set of field data that the distribution is intended to represent is illustrated in Figure 2. Figure 2 shows a curve representing a lognormal distribution that has been fit to two sets of data (one on the right and one on the left). To indicate the quality of the fit, each data set has been ordered from its lowest to its highest value, the data have then been grouped into a series of narrow ranges, and the number of values within each range plotted as a histogram on the graph. Since these then represent the

frequency distribution for the values of the data, they can be compared directly to the frequency distribution of the ideal, lognormal distribution that is represented by the curve on each graph.

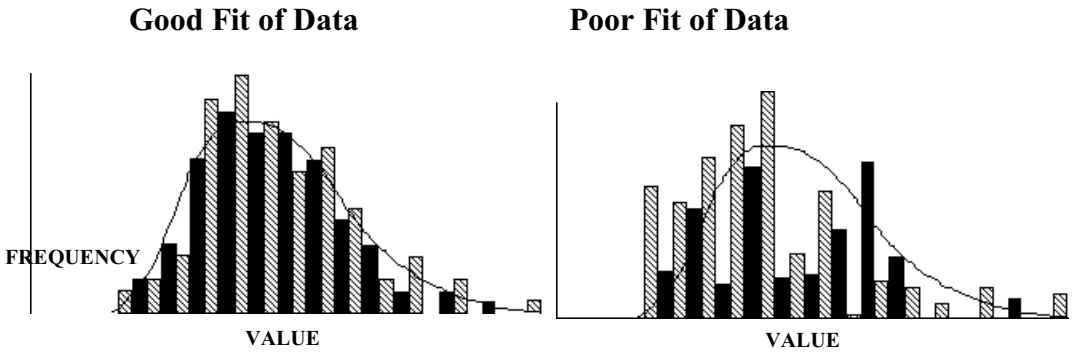

FIGURE 2 Fit of Two Data Sets to a Lognormal Distribution

The fit of the curve to the data on the left graph in Figure 2 is a good fit. This is clear from the degree of agreement between the frequencies represented by the individual columns of the histogram and the curve representing the ideal, lognormal distribution. Consequently, the properties of the ideal, lognormal distribution can be expected to corresponding to the characteristics of the data set represented on the left graph. In contrast, the data presented on the right graph in Figure 2 is poorly fit by the curve representing the lognormal distribution. Therefore, one can place little confidence in the assumption that the properties of the curve correspond to the characteristics of the data set presented in this graph.

To more formally characterize the problems that can be introduced when data are evaluated using statistical procedures (decision rules) that are not properly matched to the properties of the data, it is useful to introduce the concept of a power curve. Before employing such curves to further discussion of the effects that choice of a decision rule have on decision errors, the properties and uses of a generic (hypothetical) power curve are described briefly below.

Hypothetical Power Curve

A power curve depicts the relationship between the probability of rejecting the null hypothesis and the true condition (i.e. mean and CV) for a given decision rule. To illustrate, a hypothetical curve is presented in Figure 3. The X-axis in Figure 3 indicates one dimension of the true condition (expressed in terms of the ratio of the true mean, m, to the target value, Q). Note that it is presented on a logarithmic scale. The Y-axis is the probability of rejecting the null hypothesis (i.e. that $m < Q$). A "perfect" power curve would indicate zero probability of rejecting the null hypothesis whenever the ratio of m to Q is less than one and the probability would rise to one for all values of this ratio greater than one. This "perfect" (errorless) power curve is depicted as a dotted line in Figure 3. Note that the terminology employed in this discussion was introduced in the Background Section.

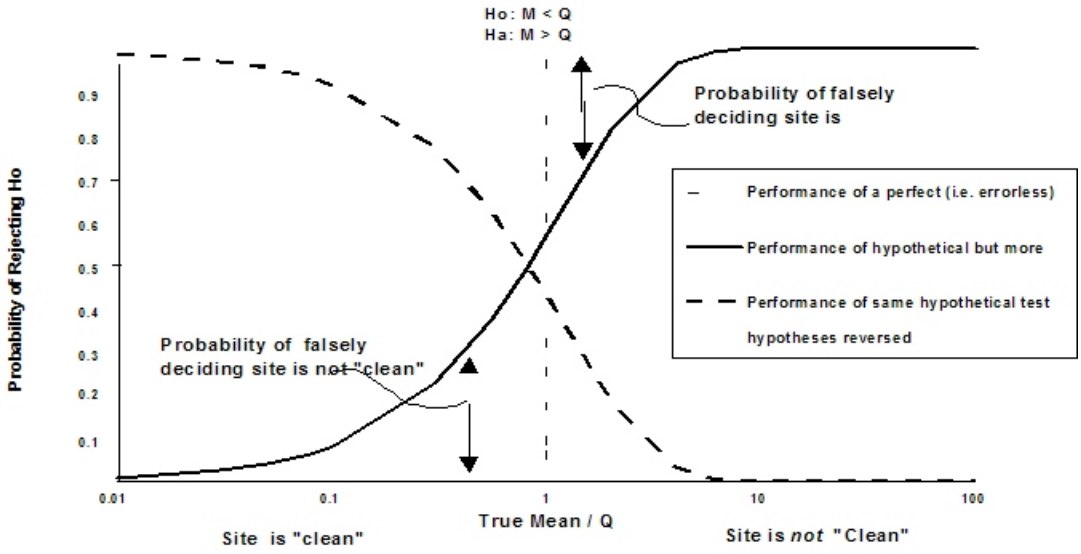

Figure 3 A Hypothetical Power Curve (Source: Berman et al. 1998a)

A more realistic power curve is depicted as the solid line in Figure 3. By comparing the solid line to the dotted line, the effects of error become apparent. To the left of one on the X-axis, the vertical distance between the solid curve and the X-axis (which is also the line representing zero probability of rejecting the null) represents the rate at which a false positive (Type I) error can occur. Thus, for this hypothetical curve, when m is one tenth of Q (i.e. at 0.1 on the X-axis), the solid curve indicates an 8% chance of *falsely* rejecting the null. This error increases as the ratio m/Q approaches one.

Once m/Q exceeds one in Figure 3, the null hypothesis should be rejected. Therefore, to the right of one on the X-axis, the vertical distance between the solid line and the line representing unit probability of rejecting the null represents the false negative (Type II) error rate. As depicted, this error rate is vanishingly small when the m is 10 times Q but increases steadily as m approaches Q from the right. The power curves presented in the remaining figures, which depict the performance of the decision rules evaluated in this paper, can be interpreted in precisely the same manner as Figure 3.

One additional feature of Figure 3 deserves mention. The dashed line in the figure represents a power curve for a hypothetical decision rule that is identical to that assumed for the solid line except that the dashed curve represents testing a reversed set of hypotheses (i.e. the alternate hypothesis for the solid line that m > Q is taken as the null for the dashed line with the old null serving as the new alternate hypothesis). This is presented to illustrate that reversing hypotheses serves precisely to transform the power curve by reflection across the line defined by 50% probability of rejecting the null. This relationship holds because the two (complimentary) decision rules would be tested using the same data set for the same true condition so that the value of the

summary statistic would be the same. Thus, all that changes is that values formerly indicating rejection of the null now indicate acceptance of the null and vice versa. The relationship between a power curve for a particular decision rule and one for the same decision rule with the hypotheses reversed was confirmed during the simulations conducted in the previous study by Berman et al. (1998a).

The Performance of Decision Rules Applied to Real Data

A set of power curves for the decision rule most commonly applied to environmental data is presented in Figure 4. This decision rule is a test of the null hypothesis that $m < Q$ in which the 95% upper confidence limit (UCL_{95}) to the mean of the data is compared to the target value (Q) to determine whether the null should be rejected. For this rule, the UCL_{95} is calculated using the Land Equation (Land 1971), which is the appropriate statistical procedure for lognormally distributed data. The various curves that are presented differ by the number of samples included in the data set. As indicated by the key at the bottom, sample sizes of 5, 10, 20 and 100 samples are represented. These curves were generated from a simulation as part of a study that was previously reported (Berman et al. 1998a).

As indicated by the curves presented in Figure 4, the chance of falsely concluding that the null hypothesis should be accepted (i.e. that the site is clean) is limited to no more than 5% by this decision rule (under the conditions represented by the figure), no matter how many samples are included in the data set. This can be seen by considering that the null hypothesis can only be false when $m \geq Q$ (which represents the parts of each curve that lie to the right of one on the X-axis) and noting that the probability of rejecting the null hypothesis when $m \geq Q$ is never less than 95% for any of the curves depicted. In fact, all of the curves pass through the point representing 95% probability of rejection when $m = Q$ (i.e., at one on the X-axis) and approach unit probability of rejection to the right.

The behavior described in the last paragraph represents the kind of statistical control that is intended by use of the UCL_{95}, as previously discussed. Because this procedure is health protective in that it limits the chance of falsely concluding that a site is clean, it has been used widely to support cleanup decisions at hazardous waste sites. As indicated below, however, the indicated statistical control is seldom achieved in practice and, even when it is achieved, it comes at a cost.

If one examines the left side of Figure 4, it is readily apparent that the decision rule depicted is not very good at limiting the probability of falsely rejecting the null hypothesis (i.e., concluding falsely that a site requires cleanup). For example, when the true mean is as little as one fifth of the target and 10 samples are employed in the analysis, there is a 30% chance of falsely concluding that cleanup is required.

The probability of falsely concluding that cleanup is required also increases as the variability or spread of the data increases. The curves depicted in Figure 4 were developed assuming that the data exhibit a Coefficient of Variation (CV) of 2. However, environmental data frequently exhibit

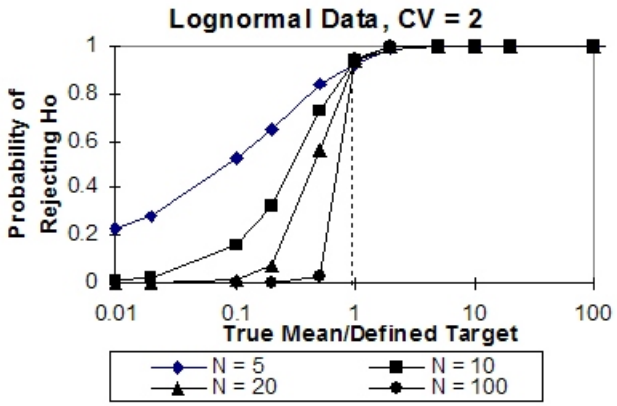

FIGURE 4 Power Curves Indicating the Probability of Rejecting the Null Hypothesis (that m < Q) Using a Comparison Between the Estimated UCL_{95} of the Mean and the Target for Lognormally Distributed Data (CV = 2)

even greater variability; CV's as large as 5 or 10 are not uncommon and can be even larger (Berman 1994). With a CV of 5, the probability of falsely concluding that cleanup is required approaches 65% for a data set containing 10 samples when the true mean is no more than one fifth of the target. Even for a data set containing 20 samples (and exhibiting a CV of 5), the chance of falsely concluding that cleanup is required is still as large as 40% when the true mean is as small as one fifth of the target.

More importantly, the error rates associated with both types of error (i.e., either falsely accepting or falsely rejecting the null) increase when the underlying assumptions concerning the characteristics of the data are not satisfied. The curves in Figure 4 were generated from a simulation in which the data sets examined were drawn from ideal, lognormal distributions. Frequently, however, environmental data do not exhibit behavior that is characteristic of a lognormal distribution. When data that are not lognormally distributed are evaluated using the decision rule described above, the true error rates that occur are greater than the error rates indicated by the ideal curves depicted in Figure 4. This is illustrated in Figure 5.

Figure 5 depicts the alpha value that is obtained for the decision rule described above when it is applied to data sets whose behavior varies from that of an ideal lognormal distribution. As indicated previously, the alpha value is the probability of rejecting the null hypothesis at the decision point (i.e., when m = Q). Also indicated previously, the alpha value for decision rule described above should be 95% when the intended statistical control is achieved.

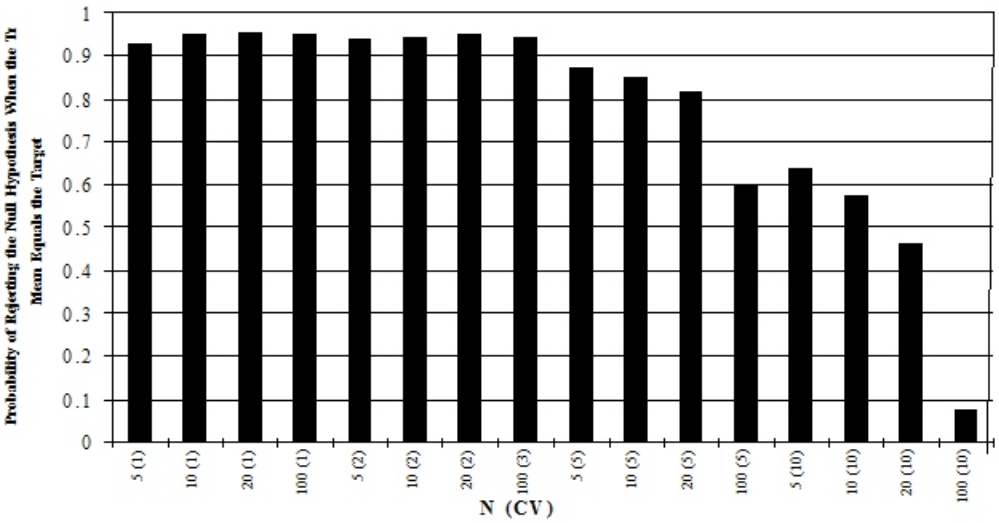

FIGURE 5 Alpha Values for a Decision Rule Involving Comparison of the UCL_{95} to the Target Applied To Data Sets Derived from a Lognormal Distribution Censored at 10% of the Target Value (Source: Berman et al. 1998b)

The data sets depicted in Figure 5 vary from ideal lognormal distributions in that they are all censored at a value set equal to 10% of the target, Q. To censor the data, any samples derived from the underlying distribution whose values are less than 0.1 x Q, were set equal to 0.05 x Q. Thus, such a data set is designed to mimic an environmental data set containing non-detects that have been adjusted by adopting a value of one-half the detection limit for each of the non-detects.

Each of the bars depicted in Figure 5 represents a censored data set containing the number of samples, "N," and exhibiting the spread (CV) that are indicated by the numbers presented on the X-axis. The Y-axis in the figure indicates the alpha value (the probability of rejecting the null hypothesis that $m < Q$ when $m = Q$) that is achieved for the data set indicated.

Although alpha values of 95% appear to be achieved for the first several of the data sets depicted in Figure 5, it is readily apparent that statistical control is lost for some of the latter data sets depicted in the figure. In fact, the alpha value for the last data set depicted in the figure (the one on the extreme right) is only about 8%. Generally, the alpha value (and therefore chance of falsely accepting the null hypothesis) increases as the spread of the data increases and as the number of samples in the data set increases. Thus, in addition to the limited ability of this decision rule to control the probability of falsely rejecting the null hypothesis, the intended control of the chance of falsely accepting the null hypothesis is also compromised when this decision rule is applied to data sets that exhibit characteristics more commonly observed among environmental data than the ideal lognormal distributions represented in Figure 4.

The performance of a variety of decision rules (in addition to the one discussed here) is reported by Berman et al. (1998a and b). Each of the decision rules evaluated are applied to simulated data sets exhibiting a broad range of characteristics that are commonly encountered among environmental data. Results from these studies indicate that, because environmental data do not typically exhibit characteristics that are sufficiently similar to the kinds of statistical distributions for which most of the available decision rules are defined, the error rates associated with application of such decision rules are generally greater than anticipated based on theory. It is further shown that decision error rates can generally be reduced by employing decision rules that are carefully matched to the characteristics of the data being evaluated. In many cases, non-traditional decision rules (which are rules that do not offer the kinds of formal statistical control that are inherent to the design of traditional rules) may sometimes be employed to reduce decision errors below that achievable using more traditional rules, as long as the non-traditional rules are carefully matched with the characteristics of the data to which they are applied.

CONCLUSIONS

As indicated above, the questionable practices identified in the Introduction Section of this paper can cause decision errors to increase when they are carelessly applied during the investigation of hazardous waste sites and the subsequent evaluation of data. In turn, this can lead both to inadequate protection of public health or to needless expenditures for cleaning up sites that otherwise pose no significant threat.

The best way to avoid the problems associated with these and other questionable practices is to carefully address data quality considerations during all stages of the site characterization process. This includes, for example, using the data quality objectives process during the planning of site investigations (so that field work is optimized for cost-effectiveness). Generally, the choice of the number of samples; the locations from which the samples are collected; and the methods used for sample collection, handling, and analysis all affect the quality of the resulting data, which (in turn) affects decision error rates. Data quality considerations should also be addressed to guide selection of appropriate tools for data evaluation that are matched both to the specific objectives of the study and to the characteristics exhibited by the data evaluated.

REFERENCES

Berman, D.W., 1994, "Data Quality, Data Quality Objectives, Good Science, and Saving Money," Workshop presented at the U.S. Environmental Protection Agency, Region 9, San Francisco, California, December 14th.

Berman, D.W., 1995, "Does Risk Assessment Work?," *Challenges and Innovations in the Management and of Hazardous Waste,* Proceedings of an Air & Waste Management Association and Waste Policy Institute Conference, Pittsburgh, Pennsylvania. Pp 493-503.

Berman, D. W.; Allen, B.C.; and Van Landingham, C.B., 1998a, "Evaluation of the Performance of Statistical Tests Used in Making Cleanup Decisions at Superfund Sites. Part 1: Choosing an Appropriate Statistical Test," *Superfund Risk Assessment in Soil Contamination Studies: Third Volume, ASTM STP 1338,* K.B. Hoddinott, Ed., American Society for Testing and Materials, Accepted.

Berman, D.W.; Allen, B.C.; and Van Landingham, C.B., 1998b, "Evaluation of the Performance of Statistical Tests Used in Making Cleanup Decisions at Superfund Sites. Part 2: Real World Implications of Using Various Decision Rules " *Superfund Risk Assessment in Soil Contamination Studies: Third Volume, ASTM STP 1338,* K.B. Hoddinott, Ed., American Society for Testing and Materials, Accepted.

Crump, K.S., 1993, "Estimating Mean Soil Concentrations When a Substantial Fraction of the Samples are Non-detects," ICF Kaiser Engineers, Ruston, Louisiana, presented to the California Department of Toxic Substances Control, Sacramento, California, March 29th.

Gilbert, R.O,.1987, *Statistical Methods for Environmental Pollution Monitoring.* Van Nostrand Reinhold, New York.

Land, C.E., 1971, "Confidence Intervals for Linear Functions of the Normal Mean and Variance," *Annals of Mathematical Statistics* 42:1187-1205.

U.S. Environmental Protection Agency, 1992, *Supplemental Guidance to RAGS: Calculating the Concentration Term,* Office of Solid Waste and Emergency Response, Publication 9285.7-081.