LA-UR- 9 8 - 3 4 2 0

Approved for public release;
distribution is unlimited.

CONF-980853--

Title: USING STATISTICS TO DETERMINE DATA ADEQUACY
FOR ENVIORNMENTAL POLICY DECISIONS (SHOOTOUT
AT THE OU-3 CORRAL)

Author(s): Elizabeth Kelly
Katherine Campbell
Dan Michael
Paul Black

Submitted to: American Statistical Association Annual Meeting
August 8-14, 1998
Dallas, TX

RECEIVED
APR 13 1999
OSTI

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

MASTER

Los Alamos

NATIONAL LABORATORY

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

USING STATISTICS TO DETERMINE DATA ADEQUACY FOR ENVIRONMENTAL POLICY DECISIONS (Shootout at the OU-3 Corral)

Elizabeth Kelly, Katherine Campbell, Los Alamos National Laboratory, Los Alamos, NM
Dan Michael, Paul Black, Neptune and Company, Los Alamos, NM

1.0 INTRODUCTION

The discipline of statistics often plays an important role in environmental policy decision-making, when decisions are, if not completely based on, at least informed by environmental data. Statistics provides guidance for the type, quantity, and quality of data required to support the policy decisions, as well as the techniques for assessing the data once it is collected. Environmental policy decisions occur at many levels, national, regional, state, and local. This paper describes the use of statistics to support policy decisions at the local level. Even at the local level, decisions can involve millions and, in some cases, billions of dollars. Additionally, local policy decisions can have ramifications for policy decisions at the state, regional and national levels.

The title of the ASA session for this paper is "Practicing Statistics at the Interface of Science and Policy." A more appropriate session title for this particular paper would replace the word *Policy* with *Politics*. In our experience, you don't get policy without politics (although the converse is not true). Our experience comes from the environmental restoration world where statistics and risk assessment are inexorably intertwined, and where the environmental data inform risk management decisions about the cleanup and restoration of waste sites and related areas contaminated with legacy wastes. The two major regulations that drive environmental restoration are the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) (also known as Superfund), and the Resource Conservation and Recovery Act (RCRA). In many areas state water quality standards and other requirements are also important drivers (see Mackenthum and Bregman 1992 for a discussion of environmental regulations).

The United States Environmental Protection Agency (USEPA) has placed the statistician right in the center of the environmental restoration work. It has done this by issuing guidance that recommends that planning for environmental data collection follows the Data Quality Objectives (DQO) Process (USEPA 1994), and evaluation of the data follows the Data Quality Assessment (DQA) Process (USEPA 1996). These processes are based on formal statistical techniques such as hypothesis testing and estimation, and explicitly link data collection to risk management decisions through specification of acceptable levels for statistical decision errors.

Environmental restoration policy decisions are often intensely political and scientifically complex, and must be made in spite of numerous uncertainties. Decisions can involve the expenditure of immense resources in terms of time, money, and people. The policy decisions often involve multiple risk managers, who are sometimes at odds with each other. For example risk managers might include regulators from multiple departments in regional EPA offices, multiple departments in state environmental agencies, and from the Department of the Interior. Local government officials and industrial or federal responsible parties are also an important part of the decision-making process. The decisions affect human and ecosystem health and often have very intense public scrutiny. In such a setting, the politics often dominate the science. Our experience is that a statistician participating at this interface should understand the politics as thoroughly as the science. Naiveté about the politics can seriously undermine the science.

The science can be even more complex than the politics. Assessing risks and providing adequate information to evaluate risk management options often requires sophisticated modeling and analysis of contaminant fate and transport by hydrologists, geochemists, geologists, toxicologists and environmental engineers. The DQO and DQA processes provide an integrating framework for the science and the politics, and provide an opportunity for the statistician to facilitate this integration.

The statistician must effectively communicate with these diverse scientists to design the environmental data collection and analysis plans. This often requires some cursory knowledge of the other scientific disciplines and the measurement techniques. The statistician must also communicate with the regulators and stakeholders to ensure that data collection designs and analysis plans address their concerns as well. Effective communication (both oral and

written) of the statistical approaches and results to this diverse audience, can be very challenging indeed. Statistical approaches must not be so complex as to appear to be "smoke and mirrors." The statistician is often working in an environment of distrust, and is always working in an area of huge uncertainties. Therefore, the statistician must make every effort to be as straightforward as possible. In our experience, statistical techniques with intuitive appeal are much more likely to be accepted. Needless to say, the formidable communication problems are a tremendous challenge for the statistician, and not every statistician will want to participate at such an interface.

This paper uses a case study to describe the political realities and challenges of "practicing statistics at the interface." Those interested in the technical components of the study should see Kelly, et al. (1995). The case study is based on an environmental restoration project we did for a Department of Energy (DOE) facility, Rocky Flats Environmental Technology Site (RFETS). Although this little drama is not a tragedy, neither does it have a happy ending. It definitely does illustrate the types of political complexities that can accompany statistics at the interface of science and policy, when

- (1) policy affects large amounts of money,
- (2) there are multiple adversarial decision-makers,
- (3) public scrutiny is intense, and
- (4) the public and the regulators do not trust the responsible party (in this case the DOE).

In our experience, these four conditions are fairly typical of environmental restoration projects. Although the paper focuses on the political challenges, the case study also illustrates the type of data commonly encountered in environmental restoration projects and some statistical techniques that can be quite useful for the data assessment phase of these studies.

2.0 BACKGROUND

RFETS is located between Denver and Boulder Colorado on the Front Range of the Rockies. The main activity at RFETS now is cleanup and decommissioning. In the past, the primary activities involved fabrication of metal components for nuclear weapons from plutonium, uranium, beryllium, and stainless steel. Activities also included chemical recovery and purification of transuranics. RFETS consists of a 385-acre industrial area surrounded by 6150 acres that comprise the buffer zone (Figure 1). The area to east of RFETS, which is down gradient and downwind, is

Figure 1. Schematic of RFETS and the Surrounding Area.

called Operable Unit (OU) 3. Operable unit or OU is an environmental restoration management term for an area that is comprised of Individual Hazardous Substance Sites (IHSSs), which are regulatory units that must be characterized and remediated if necessary. OU-3 contains four IHSSs. These IHSSs include three reservoirs and their associated drainages (outside the RFETS boundary), Great Western Reservoir and Walnut Creek drainage (IHSS 200), Standley Lake and Woman Creek drainage (IHSS 201), and Mower Reservoir and Mower ditch (IHSS 202), and the associated surface soil (IHSS 199). These reservoirs historically received waters from RFETS, but now water is diverted around them at most times. Contamination in OU-3, particularly in the reservoirs and associated drainages, resulted from accidental spills and run off from the industrial area, as well as waste water from the wastewater treatment plants. Contamination to the surface soils is largely from the 903 Pad, a former drum storage area, where an accident resulted in the windblown releases of plutonium and americium.

Our involvement with the RFETS environmental restoration (ER) program resulted from a crisis over the results of DOE audits, which brought into question the defensibility of RFETS OU-specific environmental data. These audits generally focused on data quality and integrity issues associated with the handling and documentation of individual measurements.

The DOE ER managers were interested in a DQA evaluation of the data because the DQA process focuses on the sufficiency of a data set to support decision making, rather than on the integrity of individual measurements. The DQA process would, therefore, provide another perspective on the defensibility and usability of the ER data. In addition, we explained to the DOE ER managers that in our experience many problems uncovered by datum integrity evaluations do not have a significant impact on the ability to interpret data, hence do not impact technical data usability. The primary effect of commonly noted datum integrity deficiencies is to hamper the ability to reconstruct what happened in the field and to increase the variability of the results. The DQA process evaluates whether the total variability in a data set leads to an unacceptable probability of making an incorrect decision. In addition, viewing the individual measurements in the context of the entire data set frequently reveals outliers and otherwise suspicious data. Thus, the DQA process can reveal problems with the data that the datum focused evaluations miss. While there is a role for data integrity evaluations (i.e., to explore and potentially resolve audit findings and support the assumption that data were collected in an acceptable manner), we find that the DQA process provides the most compelling, overall, data defensibility argument.

The political situation at RFETS was complex and difficult. There were adversarial relationships amongst the various groups within DOE (the auditors and the ER managers), between DOE and the various regulators, amongst the regulators, and between DOE and the RFETS contractor at the time. The DOE auditing group was pushing the DOE ER group to conduct the DQA. However, the contractor had to pay for the work out of its ER budget. This was a particularly contentious situation, because there was an impending Reduction in Force (RIF) at RFETS, which everyone knew would soon result in the loss of a thousand or more jobs. There was considerable resentment that LANL was using RFETS funding when RFETS staff were going to be RIFed. In addition, the contractor was about to lose the contract and a new contractor was soon to take over. There was concern on the part of the contractor managers, that DOE could use the problems with the data to withhold payments. Finally, the OU-3 manager was not thrilled to have his site chosen as the one to be scrutinized in the DQA. No one could blame him, but it made the job of getting the data we needed more difficult.

In addition to all of the internal politics, there had been a stop work order from EPA previously to resolve the issue of the identification of contaminants of concern (COCs). It was at this time that Dick Gilbert was consulted about comparison to background techniques and he recommended using the "Gilbert Toolbox" approach, (Gilbert 1993). The Gilbert Toolbox consists of the parametric t-test; non-parametric Wilcoxon/Gehan, quantile, and slippage tests; and, a comparison between the site data and an upper tolerance limit estimated from the background data, the "hot measurement test." If any one of these tests fails, the chemical constituent is identified as a COC. The t-test and Wilcoxon/Gehan tests look for shifts of the entire population. The Gehan, a modification of the Wilcoxon, uses a different scheme for ranking the ordered concentrations. This modification permits ranking of concentration data that contain non-detects with multiple detecting limits without making substitutions such as using 1/2 the reported detection level. The Quantile, Slippage, and Hot Measurement tests look for shifts of part of the population.

3.0 THE DATA QUALITY ASSESSMENT (DQA)

We embarked upon the DQA in the midst of all of this political turmoil, some of which we knew about, but much of which we were totally unaware. Our first step was to hold a meeting with the DOE and the contractor managers to formulate a project mission statement. To reduce the political tension and resistance, the mission statement emphasized that this would be a demonstration or pilot of the DQA process rather than an actual DQA. The steps of the DQA process include reviewing or developing the DQOs (for RFETS they needed to be developed), selecting the statistical approach, conducting the preliminary data review, verifying assumptions of the statistical approach, and drawing conclusions and making recommendations.

We learned through discussions with managers and staff that they also wanted us to evaluate the Gilbert Toolbox approach as a method for identifying COCs. They were finding that everything was showing up as a COC using this approach. Because of this, they were going to great lengths to establish what they called "weight-of-evidence" arguments to argue away COCs that were identified by the Gilbert Toolbox, but were not supposed to have been used in RFETS activities.

3.1 Risk-Based Decisions

The first step of the DQO process is to specify the decisions that the data must support. This step should involve all the risk managers; however, the decision was made to include only DOE and contractor personnel in the DQO specification. Leaving the regulators out of this stage of the process is common in our experience, and, in our experience, is a mistake. The risk-based decisions that the managers and technical staff identified were

- (1) Does the IHSS have contaminants of concern (COCs) and, if so, what are they (COC decision)?
- (2) Are there areas of surface soil that require cleanup, and, if so, where are they (cleanup decision)?

These simple decision statements belie the difficulty of actually extracting them from the decision-makers and technical staff. It never ceases to amaze us, how many data collection activities are conducted without a clear understanding of why the data are being collected and how they will be used.

These simple decision statements also belie the complexities of the policy issues that underlie them. These policy issues include the question of what constitutes data quality and how to best evaluate it. Another policy issue is what constitutes an acceptable background data set. This is often a very contentious issue with the regulators. The question of how to identify COCs is another issue that requires negotiation with the regulators. As stated previously, a stop work order over this issue resulted in an agreement to use the Gilbert Toolbox approach to identify COCs. However, because practically every chemical constituent was identified as a COC using this approach, the OU-3 staff were trying to introduce "weight-of-evidence" arguments to argue the COCs away. The regulators were not happy with these arguments. The cleanup decision required addressing the policy issue of "how clean is clean?" As we proceeded with the study, we became more and more aware that the results of our study had implications for all of these policy issues and, therefore, would be very controversial.

Because of the limitations of this paper, we present only the results for the COC decision. (The reader is referred to Kelly, et al [1995], for the interesting details of the DQA for the cleanup decision. The statistical approach was based on residual kriging and many useful techniques were developed to determine data adequacy for this decision.)

3.2 DQOs for the COC Decision

The following discussion provides the DQOs for the COC decision, developed during several meetings with the RFETS OU-3 managers and staff.

Decision Formulation:

Do the concentrations of constituents in soil or creek sediments in OU- 3 exceed those of background levels? If so, the constituents are evaluated further to determine if they are contaminants of concern (COCs).¹ If the IHSS has COCs, further investigation is required, if not, then propose no further action.

Inputs:

The variables required to support the decision are those for which data were collected. These include concentrations of Plutonium-239/240, Americium-241, Uranium in surface soil and these radionuclides and metals in creek sediments.

Boundaries:

Surface soil (1/4" to 2" in depth) within areas that are (1) within the geographic boundaries of OU-3, and (2) are in areas considered to be representative of background for the region (unaffected by Rocky Flats or other potential sources).

Creek sediments (represented by grab samples) from creeks and ditches within the geographic boundaries of OU-3, and corresponding sediments from sources considered to be representative of background

Decision Rule:

The decision rule for this decision was established previously through an agreement to conduct appropriate background comparisons derived from the five tests described in the Gilbert Toolbox for each constituent. If any appropriate test fails for a constituent, then that constituent is evaluated further to determine if it is a COC.

Limits on Decision Errors:

The OU-3 staff preferred not to set power requirements for this demonstration DQA. Instead, they requested power curves for OU-3 data showing the level of power achieved for various hypothetical differences between the site and background data. The power curves were based on OU-3 sample sizes, Coefficients of Variation (CV) and RFETS background distributions. (Note that the regulators specified a Type I error of 20% and a Type II error of 10% for a 20% shift in location.)

3.3 Preliminary Data Review

The preliminary data review had the usual logistical complexities, very large databases that needed to be acquired, cleaned and integrated. Although simple in principal, this part of the study took longer and was more frustrating than we anticipated (our experience indicates that this is often the case). The background data sets had from 46 to 58 measurements per chemical constituent. Great Western Reservoir and Walnut Creek drainage (IHSS 200) had eight, Standley Lake and Woman Creek drainage (IHSS 201) had fourteen, and Mower Reservoir and Mower ditch (IHSS 202) had four, for a combined sample size of 26. The surface soil data for IHSS 199 contained approximately 60 observations while the background data (Rock Creek) contained 20 observations. Several of the data sets had non-detects. Table 1 illustrates the varying numbers of non-detects for the three radionuclides and four of the metals in the creek sediment data.

Table 1. Examples of Non Detects in the Data.

CHEMICAL CONSTITUENT	OU-3 Sediment Non Detects Sample size = 26	Background Non Detects Sample size = 46 to 58
Arsenic	0	6
Beryllium	0	29
Silver	10	52
Molybdenum	14	30
Pu - 239/241	3	5
Am - 241	2	7
U-238	0	0

Simple tools were used in the data review, including box plots and probability plots. These tools revealed several obvious statistical outliers in the background data sets. Outliers can dramatically affect the slippage test and the UTL calculations. The quantile and t-test are somewhat less sensitive, but certainly can be affected by outliers. The Wilcoxon / Gehan test is fairly robust to outliers. Statistical outliers were removed from background data sets. Figure 2 provides a clear example of the presence of an outlier in the arsenic sediments background data. The data point in question at 17.3 mg/kg is far in excess of the remainder of the background data, and far exceeds published ranges for background arsenic concentrations. The second greatest arsenic concentration, at approximately 7 mg/kg, is also, potentially, an outlier. However, there was insufficient geological or other supporting information to be able to make a determination for that observation, so it was included in the ensuing data analyses. The data review revealed that many of the background data sets contained statistical outliers.

Figure 2. Arsenic Boxplots Showing Statistical Outlier in Sediment Background Data

Probability plots were used to study the distributions of the background data. Figure 3 shows two probability plots for Arsenic (with the outlier removed). The first plot is for the untransformed data (linear scale), and the second for logarithmically transformed data (logarithmic scale). In neither of these plots do the ordered data fall along a straight line. (Points represented by one-half the reported detection limit are shown as squares in these plots). On the linear scale the data curve upward, as would lognormal data. On the logarithmic scale they fall off at the lower end, although the upper half of the data are reasonably linear. Although many observations are reported as detected as low as 0.5 mg/kg, the associated measurement errors are relatively large near the detection limit, making these observations more variable and emphasizing interlaboratory differences (this situation is also observed at LANL). In such cases we do not worry too much about the lack of fit of the lower end of the distribution. Thus, arsenic data were used as the model for the lightly censored lognormal case for the power simulations. Figure 3 also shows the UTLs for arsenic data, drawn as solid lines on the boxplot graphs.

Figure 3. Probability plots and UTLs for Arsenic Concentrations

Figure 4 shows the probability plots and UTLs for beryllium. The beryllium data appear to be normal except for the lower end, which is almost entirely non-detects. These data are so highly censored that it is impossible to definitively determine whether the underlying distribution is more nearly normal or lognormal. Beryllium was used as the model for heavily censored normal case for the power simulations

The distributions for the Molybdenum and Uranium background data were less ambiguous. Molybdenum (once problems with detection limits were resolved) provided the basis for the heavily censored lognormal case and the uranium data provided the basis for the lightly censored normal case for the power simulations

Figure 4. Probability plots and UTLs for Beryllium Concentrations

3.4 Verifying Assumptions

In an actual DQA this step would include determining if the data provide the specified significance level and power. However, for this demonstration, significance and power were not specified, instead power curves were to be evaluated. There was concern that the significance level for the Gilbert Toolbox was too high, since almost every chemical constituent was identified as a COC using the Gilbert Toolbox. Even the major elements such as iron, calcium, sodium, potassium and magnesium were identified as COCs. Very few (if any) of the major elements should have been different from background.

3.4.1 Power Simulations

Simulations based on the background data sets were used to develop power curves for the various chemicals. The simulations captured eight situations seen in the OU-3 site and background data:

- Both background and site populations were either normal or lognormal with lognormal censoring.
- Background and site distributions had either light censoring (typically 10% or less in the background population) or heavy censoring (typically 50% or more in the background population). The censoring mechanism was simulated as a lognormal distribution truncated at some low but positive value c . Every simulated background value below this truncation level was reported as a non-detect with detection limit c . Higher values were reported as non-detect if and only if the simulated detection limit for that "observation" exceeded the simulated background value. The postulated lognormal censoring models produced censoring similar to that observed in the actual data.
- Site populations shifted to the right either entirely or partially. For the former case, the population model was obtained by multiplying the background population mean by factors from 1 (i.e., no shift) to 2, and preserving its coefficient of variation (CV) (so its standard deviation was also increased by the same factor.) A partial shift is a combination of the observations (on the average) from a population that has been shifted, as above, by a factor of 2, and the remaining from the background population, where the partial proportion varies from 0 (no shift) to 1 (entire shift). The CVs were based on the various background data sets.

Power calculations were run for one background sample size (50), for 6, 12 and 18 site samples, and for three significance levels (0.05, 0.1, and 0.2). These sizes approximate the sample sizes observed in the OU-3 background and site data. In order to evaluate the performance of UTL comparisons, the (0.99,0.95) Figure 5 illustrates the power curve calculations for the individual tests of the Gilbert Toolbox and the tandem test (one of four) for a site sample size of 12, assuming partially shifted alternatives. The curves are not smooth because they are estimated using only 100 simulations. UTLs were calculated for four sediment data sets, approximating the four background types. The hot measurement test was evaluated for both the normal and lognormal UTLs for each case.

Results. The power simulations showed that the significance level for the tandem test (the Gilbert Toolbox without the hot measurement test) varied from 0.06 to 0.11 when the individual tests were run at the .05 significance level. For a relative shift of 20%, whether achieved by a 20% shift of the entire site population or a shift of 20% of the site population by a factor of 2, the power of the tandem tests ranges from about 20% for 6 site samples up to 30-40% for 18. Power is better for the lightly censored cases than the heavily censored ones. Except in the heavily censored lognormal case, power against a relative shift of 50% approaches 80% for the largest site sample size (18). These results suggest that demanding 90% power against alternatives much less than a shift by a factor of 2 will require very large sample sizes, although 80% power for a relative shift of 50% could be achieved with a sample size of 20 or so.

It should be noted that all of the cases simulated had relatively high coefficients of variance (in the range of 0.6 to 0.85), so that a shift of 20% in the mean is a shift of only .25 to .3 standard deviations. Even a shift by a factor of 2 (with a coefficient of variation of about 0.6) is less than two standard deviations of the background distribution, further diluted by the increase in variance in the shifted distribution. Increased power for smaller samples would be observed in cases with smaller coefficients of variance, but the background creek sediment data sets that were the bases for these simulations exhibited high coefficients of variation.

Ninety percent power at a relative shift of 50% can also be obtained for a sample size of 18, if the component tests are run at a nominal 0.1 significance level, at the expense of false positive rates of about 20% for the tandem tests. But even if the tests are run at a nominal 0.2 significance level (with false positive rates of 30% in most cases), the power of the tandem test against a shift of 20% is still only in the 50-60% range for 18 samples.²

A hot measurement test contributes little to the package, at the expense of many false positives for the lognormal case, if the normal UTL happens to be used. However, it is quite successful in detecting partial shifts, and could be recommended for larger sample sizes using a "one above lognormal UTL or two above normal UTL" rule.³

Figure 5 Power Curves for Sample Size 12 with Partially Shifted Alternatives

3.4.2 Appropriateness of Background Data

The reliability of the statistical tests is not only a function of sample size, fraction of non-detects, and other quantities that determine statistical power, but also, and even more critically, on the original premise that the background sample is indeed appropriate. In general, demonstrating that data correctly represent the population that they claim to represent, whether it is background or site data, is the most difficult part of the data-based

environmental decision-making. Statistics alone do not replace the need for scientific evaluation and judgement, but, as shown in this study, statistics can help to identify possible problems.

Figure 6 shows that a boxplot comparison of molybdenum background data collected in different years revealed significant differences between background data collected in 1989 and in the 1990s. This difference appears to be a result of very elevated detection limits in 1989.

Figure 6. Boxplot Comparison of Molybdenum Background Concentrations Collected in 1989 and in 1990s

Figure 7 shows that a boxplot comparison of background and site data for plutonium raises questions about the appropriateness of the plutonium background data, since the site data had lower levels than the background data

Figure 7. Boxplot Comparison of Plutonium Background and Site Data

Table 2 shows that one possible reason that the major elements were all identified as COCs by the Gilbert Toolbox test is that the background data set was geochemically different than the site data. This table shows that levels of iron in the site data were considerably higher than in the background data. Several chemicals are known to be highly correlated with iron, including arsenic, one of the COCs that RFETS was trying to eliminate using “weight of evidence” arguments. Table 2 shows that the expected levels of arsenic in the site data, based on the correlation between iron and arsenic in the background data (0.7), were not significantly different than what would be expected from background with comparable levels of iron. It is interesting that the RFETS approach was to blame the statistical test for too many COCs (false positives), not the data. As a consequence, they spent considerable resources developing “weight of evidence” arguments to try to eliminate arsenic as a COC.

Table 2 Expected Levels of Arsenic Given Levels of Iron

LOCATION	OBSERVED MEAN OF IRON (mg/kg)	EXPECTED MEAN OF ARSENIC GIVEN IRON (mg/kg)	OBSERVED MEAN OF ARSENIC (mg/kg)
IHSS 200	25,816	7.05	5.31
IHSS 201	15,398	4.20	4.76
IHSS 202	19,200	5.24	4.88
Background	8,853	2.41	2.41

3.5 Drawing Conclusions and Making Recommendations (Shootout At The OU-3 Corral)

Five months after beginning the study, we delivered a three-volume report to DOE and RFETS managers and I gave a final briefing presenting the results and our recommendations. The results of the COC DQA demonstration included:

- Site data generally usable (for reasonable power and significance specifications)
- Background data has issues that need to be addressed – outliers, detection limits, appropriateness
- Gilbert Toolbox Tandem Test (1 in 4) useful for COC specification
 - Significance level around double individual levels
 - Power reasonable (for sample sizes of 12 or more) (although not that different than t-test in most cases)
 - UTL adds nothing and currently has problems because of background outliers and detection limit problems

The recommendations included:

- Negotiate reasonable power and significance requirements with regulators.
- Have statisticians and geochemists work together to resolve background issues.
- Use tandem test with “one above lognormal UTL or two above normal UTL” rule.

It is said that a picture is worth a thousand words. To understand the enthusiasm with which these results and recommendations were received, the reader should picture the infamous shootout at the OK corral. Without going into details about who was shooting at whom, let us just point out that this was not a case of “don’t shoot the messenger.” The shootout at the OU-3 corral occurred because of all of the politics previously mentioned and, because the good news as to the usability of the site data was far overshadowed by the bad news about the problems with the background data.

4.0 WHY PRACTICE AT THE INTERFACE?

Given the difficulties described in this paper, why would a nice statistician like you (or us) want to get involved at such an interface of science and policy/politics? A perfectly valid question, however, we believe there are incentives. Below are some that come readily to mind.

- You will never be bored.
- There is a lot of action and you are a key player (of course, you have to like the action).
- Becoming knowledgeable about the diverse scientific disciplines is both challenging and interesting.
- Working in a team environment can be quite stimulating and a lot of fun (always someone to debate over a cup of coffee or a beer).
- Developing the required communication skills is challenging and, sometimes, quite rewarding.
- If you are looking for a growth experience, this work provides ample opportunity.

5.0 REFERENCES

Gilbert 1993. Richard O. Gilbert, “Recommended process for implementation by Rocky Flats Plant (RFP) to compare environmental restoration site analytical results obtained in operable units (OU) to background concentrations,” July 30, 1993.

Kelly et al. 1995. E. Kelly, K. Campbell, D. Michael, P. Black, “Environmental Restoration Data Evaluation Pilot Project - Volume 1: Managers Summary, Volume 2: Data Quality Assessment / Proposed Framework and Demonstration of the DQA Process At RFETS,” RF/ER-95-0074.UN, June 1995.

Mackenthun and Bregman 1992. K. M. Mackenthun and J. I. Bregman, *Environmental Regulations Handbook*, Lewis Publishers, Inc. Chelsea, MI, 1992.

USEPA 1994. Guidance for the Data Quality Objectives Process, USEPA QA/G4, September 1994.

USEPA 1996. Guidance for Data Quality Assessment Process, USEPA QA/G-9, QA96 Version. July 1996.

¹ The “further evaluation” consisted of elaborate “weight-of-evidence” arguments that were based on other background studies, not necessarily from the region. At the time we were involved, the regulators did not accept these arguments.

² An interesting statistical result is that for the situations simulated here, the t-test retains its optimality properties for distinguishing between two distributions, even when most of the assumptions underlying that optimality are violated. Using the t-test in tandem with the Gehan, quantile and slippage tests results in rejecting the null hypothesis in only a slightly larger fraction of cases for a given alternative. See Kelly et al (1995) for more discussion of the statistical results for the various tests.

³ An interesting aside is that the hot measurement test is the one preferred by field investigators and regulators alike at both RFETS and LANL. They prefer a single number for quick comparisons, rather than running multiple tests.