Statistical Experts’ Workshop

Seattle Public Library, Oct. 7, 2008

Sponsored by:

Regional Sediment Evaluation Team
DMMP Dioxin Workgroup
EPA Region 10 Superfund

Workshop Report

Prepared by:

Teresa Michelsen
Avocet Consulting
Olympia, WA

Introduction

The Regional Sediment Evaluation Team (RSET) partners, including OR DEQ, WA DOE, EPA, NMFS, USFW, and the Corps of Engineers, are working together to develop a Sediment Evaluation Framework (SEF) for use in Oregon, Washington, and Idaho. The SEF includes procedures for conducting a bioaccumulation assessment of dredged material, including tissue and sediment concentrations protective of aquatic life, and human and wildlife consumption of fish and shellfish. Similarly, the Puget Sound DMMP is involved in determining appropriate bioaccumulation-based Screening Levels for sediments for use in the open-water disposal site program for dioxins/furans and coplanar PCBs. Finally, EPA Region 10, WA Dept. of Ecology, and OR Dept. of Environmental Quality are encountering bioaccumulative compounds at numerous State and Federal Superfund sites in the Pacific Northwest.

In each of these programs, many of the risk-based guidelines for bioaccumulative compounds in sediment, as calculated using standard EPA risk assessment methodology, are below either natural or globally-distributed background concentrations. In this case, a comparison of project sediments to background concentrations may be conducted for a variety of dredging, disposal, and cleanup-related objectives. In doing so, numerous questions arise of a statistical nature, such as:

- Whether to use single-sample or population comparisons
- What “bright line” values are available to characterize background concentrations and what are their pros and cons
- What population-based comparisons are available and what are their pros and cons
- How these comparisons are affected by sample size and distribution
- How outlier tests should be used in defining background distributions
- Whether there are multi-part or multi-step comparisons that could be used
- How non-detects should be addressed, both for single chemicals and summed suites of analytes
- What null and alternate hypotheses should be used
- What statistical software is appropriate for the recommended approaches

The various agencies and workgroups above decided to convene a workshop with expert statisticians to provide their best professional experience and advice on how to conduct these comparisons.

Process

Both the RSET Bioaccumulation Subcommittee and the DMMP Dioxin Workgroup were requested to propose experts for the workshop, and also to frame the questions to be answered. The scope of the workshop included comparisons to background in the context of both dredging and cleanup projects, but did not include the development of background data sets or policy questions related to the use of the data or the establishment of DQOs.

From the experts nominated, eight were invited and four were able to attend, including:

- **Loveday Conquest**, Associate Director, Aquatic & Fishery Sciences, UW
  Professor, Center for Quantitative Science
- **Dennis Helsel**, Practical Stats
  Author of "Non-Detects and Data Analysis" and "Statistical Methods in Water Resources, teaches classes on non-detects and multivariate analysis
- **Lorraine Read,** TerraStat Consulting Group
  Consults on statistics on many NW projects and has taught courses such as Environmental Statistics for Site Managers.
- **Anita Singh,** Lockheed Martin Environmental Services
  Developer of ProUCL and multiple EPA guidance documents

Based on these efforts and a large number of submitted comments and questions from both workgroups, background information, an example data set, and a set of questions were prepared and submitted to the experts in mid-September. This information is included as Attachment A, and was written largely by John Wakeman (Seattle District COE) and Teresa Michelsen (Avocet Consulting).

The expert panelists reviewed the questions and provided the moderator (Teresa Michelsen) with initial thoughts, comments, and responses. These were compiled into a single set of notes for use at the workshop by both the panelists and the audience. Based on the initial responses, some questions were dropped from the workshop and others were re-ordered and streamlined, as it became clear that there was consensus on certain issues ahead of time. The following report addresses the topics as they were discussed at the workshop, and the two questions that were not addressed at the workshop are discussed as part of question 10.

Each section describes the ultimate recommendations of the expert panel, who were generally in consensus on all points. Where there was more than one approach suggested but acceptable to all, those are indicated as alternatives. Because the workshop was ahead of schedule most of the day, it was possible for the audience to participate and ask questions, adding depth to the discussion. A list of attendees is provided as Attachment B.

Workshop Goal

The overall question for the day was how to perform a comparison of project sediments (either sample-by-sample or as a distribution) to a background distribution or threshold to determine whether the project samples exceed the background population concentrations.

The following sections discuss the individual topics and questions within the topics. In each case, the question as summarized for the workshop is listed first, followed by the condensed notes from the initial expert responses. Following that, the consensus of the discussion is stated.

Question 1. What are the statistical advantages or disadvantages of using a comparison of **project and background distributions** vs. **comparison of individual test samples to a "bright-line" threshold** based on the background distribution? Does this answer depend on the size of the background data set and/or its distribution? If so, how?

What is the question: need to know about individual samples, or overall distributions?

- ⇨ Bioaccumulation is more of a population (area) issue than benthic toxicity, which suggests that the statistical test should address the question of 'chronic' or average exposure using the entire project distribution.
- ⇨ However, if project data ≠ background, need to know which samples/areas are above background (to designate as unsuitable for open water disposal or for further cleanup). This indicates the need to identify individual sample violations of some acute bBackground threshold.

Project Distribution to Background Distribution Comparison:

- ⇨ **Pro:** Incorporates uncertainty in both project and background distributions
- ⇨ **Pro:** Controls site-wide false positive error rates, whereas individual sample comparisons do not)
- ⇨ **Con:** Distributions being compared may not be the same shape
- ⇨ **Con:** Requires more project samples
- ⇨ **Con:** Gives no information on individual samples
- ⇨ Can test equality of the means or medians (best for total mass of contaminant)
- ⇨ Can test what proportion of samples are below an upper threshold of background

Individual Project Sample to Background Bright-Line Comparisons:

- ⇨ **Pro:** Better reflects regulatory decisions (need for info on individual DMMUs, specific site stations or areas)
- ⇨ **Pro:** No minimum data set size for project data
- ⇨ **Pro:** Do not need to determine distribution for project data
- ⇨ **Con:** Gives little information on uncertainty or error rates for project data
- ⇨ Can focus on either the upper end or the mean of the background distribution
- ⇨ Best if there are concerns about individual high samples

Consensus Points:

- It will be necessary to calculate a bright-line upper threshold from the background distribution, as there are many dredging projects (and some initial investigation data sets) that will have too few data for population comparisons.

- It would also be valuable to make the background data set from which the bright line is calculated available for population comparisons for projects (such as larger cleanup sites) that do have enough project data, as this is considered a preferable approach.

- A minimum of 10 samples in the project data set is recommended for population comparisons performed using single or two sample hypotheses. However, whenever possible, it is suggested to determine sample sizes based upon pre-established DQOs to estimate bright-line or background threshold values or to compare background versus project concentration distributions.

Question 2. If the agencies chose to use a bright-line threshold for the background distribution (UCL, UPL, UTL, upper percentile, etc.), what alternatives are available, and what are their pros and cons? Does the recommended threshold depend on the size of the background data set and/or its distribution?

- ⇨ **UCL** is an upper confidence limit on the mean, and will produce large numbers of false positives if compared to individual project samples, therefore **not recommended** for that purpose. OK if comparing to a statistic on the mean of the project distribution.

- ⇨ For upper tail of background distribution, **UPL vs. UTL?** Both seem OK, which would be recommended and why?

  UPL = Upper prediction limit - the upper bound of the background dataset, below which k future observations drawn from the same population are expected to fall with specified probability.

  UTL = Upper tolerance limit - a confidence limit on a percentile of the underlying population.

- ⇨ What should be used if significant non-detects are expected?

- ⇨ What about for small background data sets?

Consensus Points:

- UTLs were recommended for calculating bright-line thresholds. They are easier to understand and explain and are based only on the background distribution. UPLs are project-specific and depend on the number of future observations (k) that are expected to fall below the UPL (assuming they are from the same population). As k increases, the calculated UPL must also increase.

- UTLs come in two forms:
  - **Content UTLs** contain ***at least*** X% of the population with Y% confidence
  - **Expectation UTLs** contain ***on average*** X% of the population with Y% confidence

  Both of these options are available in ProUCL.

- The agencies will need to make some decisions regarding the specific percentile (X) and confidence (Y) on the percentile to use (policy decision). However, the higher the percentile, the more the result will be affected by extreme values in the background distribution. Note that **content** UTLs have been used for regulatory decision making in San Francisco using 95% confidence on the 85th percentile for sediments (e.g., Gandesbery et al. 1998), on the 90th percentile for bioassay results (e.g., Germano &

Associates 2004; or Hunt et al. 1998), and by USEPA using 95% confidence on the 90th percentile (RCRA Guidance, 1998 addendum).

- Non-parametric methods (e.g., bootstrap and Kaplan-Meier methods) are available in ProUCL to calculate UTLs and should be used if non-detects form a significant part of the data set and/or if the background distribution is not normal (more on that later).

- If the background data set is small, a non-parametric UTL will likely be the maximum value, which may actually correspond to a lower percentile of the distribution than the one desired.

- To estimate a bright-line threshold (by any of the methods above), every effort should be made to collect enough data based upon DQOs. At a minimum, at least 10 observations should be made available to estimate threshold values.

Question 3. If the agencies chose to use a population-based comparison, how does the type of distribution(s) affect the selection of a test to compare the project data with the background data? Which parametric or nonparametric tests should be used for various distributions? Is a flow diagram needed to guide project proponents and staff in making these decisions? If so, what would that look like?

Parametric tests:

- ⇨ **Con:** Background and project distributions may not be the same shape
- ⇨ **Con:** Project distributions may be hard to classify
- ⇨ **Con:** Don't handle non-detects well, which may be high in both data sets for some chemicals
- ⇨ Recommendation - do not transform data to achieve normality

Non-Parametric tests:

- ⇨ **Pro:** rank-based tests are not affected by the cons listed above for parametric tests.
- ⇨ **Con:** rank-based tests do not contain information about magnitude of the difference in the original measurement scale.
- ⇨ **Mann-Whitney-Wilcoxon** and the associated **Gehan Test** (a generalized Wilcoxon test for use when the data contain multiple detection limits) – a 2-sample test of location (addresses the question: does one distribution tend to have higher values than the other?).
- ⇨ **Quantile Test –** 2-sample rank test to check for a shift in the tail of one of the distributions.
- ⇨ **Kolmogorov-Smirnov** – a comparison of the observed cumulative distribution functions (addresses the question: do the two distributions look the same in shape, spread, and location?)

Consensus Points:

- Current thinking is that substitutions for detection limits should generally never be done, and transformations to achieve normality also should not be done, because important information about the original data set is lost. Because of these two factors, traditional parametric tests are not going to be useful in most cases, especially when data sets consist of non-detects. In the past, these were used because of limited computing power and unavailability of inexpensive software.

- Good non-parametric tests exist and can easily be performed with current computing capabilities that do not require substitutions or transformations. These tests are recommended in all cases, even if data are approximately normally distributed. Consistency of methods across all sites and times is valuable, and little if any gain in power is achieved with parametric tests unless data are exactly, not just approximately, normally distributed. More guidance on these topics can be found in the ProUCL 4.0 Technical Guide.

- The experts selected two tests for a nonparametric population comparison, depending on whether the data contain multiple differing detection limits or not:

  - **Wilcoxon-Mann-Whitney** (WMW) test for data without non-detects or a single detection limit
  - **Gehan's test** (generalized Wilcoxon) for data with multiple detection limits

  Both of these tests are rank-based and they compare the relative distribution functions to determine whether the values in one group are generally larger than the values in the other group, and do not require substitutions. Both are available in ProUCL.

- A population comparison is not recommended for project data sets with few data, because these tests are relatively insensitive to high values at the upper tail with small data sets.

Comparing populations using these tests alone does not describe the magnitude of the shift or differences in the upper tails, and therefore additional evaluations may be appropriate on a project-specific basis. The Quantile test (also in ProUCL 4.0) is often used to determine if the upper tail of project distribution is higher than the background distribution. It is suggested to use the Quantile test in tandem with the Wilcoxon-Mann-Whitney Test.

<u>Question 4.</u> How does the presence and/or percentage of non-detects in the data set affect the selection of a statistical test? Should this be factored into the flowchart?

- ⇨ Use non-parametric methods, as it is difficult to justify the use of GOF tests on data sets with non-detects
- ⇨ Avoid substitutions (see below)

Consensus Points:

- By this point in the workshop, this question had largely been answered. The specific percentage of non-detects was not considered important, but there was a strong emphasis on not conducting substitutions and instead using nonparametric tests to handle this situation.

Question 5. What alternatives are available for addressing non-detects (in either distribution), what are their pros and cons, and which would you recommend for individual compounds?

- Does the answer depend on the percentage of non-detects in the data set, and if so, how?
- What is the effect upon the robustness of the testing of replacing non-detected values with reporting limits, ½ reporting limits, 0, or a Regression-on-Order (ROS) statistic while performing a parametric test?

⇨ Cannot make distributional assumptions with high NDs
⇨ Avoid substitutions
⇨ ROS methods?
⇨ Kaplan-Meier and bootstrap methods

Consensus Points:

- Substitutions of any kind create “invasive data” that originates from the laboratory (because it is relative to the DLs) rather than from the environment. Therefore, given that there are now good alternatives, this should essentially never be done.

- Non-parametric tests work well for establishing bright lines for background distributions with non-detects regardless of the percentage of non-detects, because they generally count down from the top of the distribution and we are usually interested in upper percentiles. **Kaplan-Meier** is the approach specifically recommended (and can handle both single and multiple detection limit situations).

- In general, ROS approaches are not necessary, since there are non-parametric tests that do not need this information, and the experts agree that in general, Kaplan-Meier will give better results. However, non-parametric tests will not go beyond the bounds of the actual data, and therefore there may be occasions on which an ROS approach is useful:

  - Attempts to calculate an upper percentile in a small data set will go no higher than the actual maximum value, which may be a lower percentile than is desired

  - If there are enough non-detects that the upper statistic being calculated is within the non-detects, then ROS will provide a solution.

Question 6. Using the data set provided as an example, how should non-detects within summed classes of compounds be addressed? For instance, congeners may be detected at various frequencies, from 0-100%. However, because they are all typically added together on a station-by-station basis to calculate an overall TEQ, a single type of non-detect substitution is typically employed (e.g., ½ the reporting limit). Is this appropriate? If so, what is the best substitution to use, and if not, what alternative approaches are available?

- ⇨ Substitutions not recommended
- ⇨ Substituting 0 and the DLs will provide a range within the true value lies, though the range may be too wide to be useful.
- ⇨ Use Kaplan-Meier (to estimate mean vector and covariance matrix)
- ⇨ Use multivariate methods (# samples > # compounds)
  - **Pro:** Creates a prediction (tolerance) ellipsoid
  - **Pro:** Scout 2008 has these methods
  - **Con:** requires too many samples for some project distributions

Consensus Points:

- Even for summed compound classes with varying degrees of NDs among the congeners, it may not be necessary or appropriate to conduct substitutions. There are at least two approaches available for handling this situation, outlined below.

- For multivariate data sets with and without non-detect observations, appropriate methods can be used to compute multivariate background threshold values. These methods require that the number of samples (observations) is larger than the number of variables (analytes). In summary, the process involves an n x p background data matrix (where n = number of samples and p = number of variables or analytes), and computing a mean vector and a covariance matrix based on the background data set. Multivariate individual and simultaneous thresholds are computed using a chi-square or scaled beta distribution.

  Control chart-type index plots (similar to a univariate control chart), prediction ellipsoids (corresponding to a UPL in the univariate case), and tolerance ellipsoids (corresponding to a UTL in the univariate case) drawn at background threshold levels can be used to identify project samples that may not come from the background population. Specifically, on the control chart-type index plot, project MDs lying above the simultaneous background threshold can be considered as representing observations not coming from the background population. The details of this procedure can be found in the Scout 2008 User Guide.

- Multivariate control chart-type index plots can identify observations with a shift in the mean as well as observations that may not comply with the covariance structure displayed by the background data set. One important point to note that is just like univariate background data sets, multivariate background data sets should also be free of outliers. Scout 2008 has several multivariate robust outlier identification methods (e.g., PROP, Huber, and Biweight influence function-based methods, MCD method) that may be used to identify multivariate outliers in a background data set.

- The question then arose, was there any way to calculate a sum (e.g., TEQs) for a single sample without substitutions. Dr. Helsel presented one possible approach, which is newly developed (will be presented at SETAC in Nov. and published thereafter). Very simply:

  If the mean of n observations is the sum divided by n, then the sum = mean * n.

  The mean in this case can be determined using the Kaplan-Meier technique without substitution, and simply multiplied by n to arrive at a TEQ. Each value would have to be weighted by its TEF first. Peer review and testing of this method is still underway, but early tests are positive.

<u>Question 7</u>. Given the answers to the questions above, what specific form of the hypothesis would be recommended, including a) null hypothesis, b) alternate hypothesis, and c) α and β levels.

- ⇨ Form 1, no S
- ⇨ Null hypothesis = site/project data ≤ background
- ⇨ Alternate hypothesis = site/project data > background

Consensus Points:

- **Form 1 hypotheses** assume compliance. A small data set has difficulty rejecting the null hypothesis because there is little evidence. Therefore Form 1 usually is accompanied by a minimum sample size so that, if there is difference from background, it might be detected.

- **Form 2 hypotheses** assume noncompliance. The burden of showing compliance is placed on those regulated, and there is therefore an innate incentive to collect sufficient data to reject the null hypothesis and conclude compliance. No separate requirement of minimum sample size is therefore required, but the assumption of noncompliance may not be acceptable.

The form of the hypotheses, or alpha and beta, were not directly discussed at the workshop, as these are primarily policy decisions. The following are some examples.

Comparison of distributions using rank-based Mann-Whitney type test. Alpha is fixed (traditionally at 0.05 or 0.10); beta (and power) are a function of alpha and sample size and will be the same regardless of the variance of the data because ranks are used.

- **Form 1** (Assumes Compliance)
  - Null: Pr(Site < Background) ≥ 0.5
  - Alternative: Pr(Site < Background) < 0.5

- **Form 2** (Assumes Non-compliance)
    - Null: Pr(Site < Background) $\leq$ 0.5
    - Alternative: Pr(Site < Background) > 0.5

Comparison of individual sample to bright line. Policy decisions regarding the confidence level and the content coverage of a UTL need to be made. An example is provided below.

- Definitions:
    - X ~ Background distribution
    - Background Threshold Value (BTV) is the 95% confidence limit on the 85th percentile
    - X_{85} is the true 85th percentile of the background distribution
    - Y is an observation from a project site

- Pr(X_{85} < BTV) = 0.95. Then it follows that if (Y < BTV), this implies that Y $\leq$ X_{85}. From this we infer that the new observation Y does not fall into the upper tail of the background distribution and is therefore not too dissimilar from background.

- The power (i.e., the probability of concluding that Y is not from background when it really isn't) cannot be calculated without specifying the alternative distribution (so power as a function of sample size and various differences).

Question 8. What is the role of outlier tests, particularly in establishing a background data set, and which method would be most appropriate?

⇨ Very important to ID outliers in background data sets
⇨ Must determine whether it is possible that extreme values are either:
1) from a different distribution
2) an extreme tail of the distribution
⇨ Avoid mixing more than one population; stratify if necessary
⇨ Requires knowledge or assumption of the background distribution
⇨ A large amount of data is helpful
⇨ Use graphical methods

Consensus Points:

- It was agreed that the approach originally laid out in TerraStat's memo of August 17, 2007, was a good starting point, with some additional detail on the specific tests. In summary:

    1. Create box plots (for individual data points or MDs) and use a cutoff value at the upper end to identify potential outliers (e.g., 1.5 x the inter-quartile range above

the 3rd quartile). Graphical methods are in general encouraged for identifying outliers.

2. Examine the data quality for these values and/or whether or not there is a reason to believe that there might be non-background sources present. If so, eliminate. If not, subject to further tests.

3. In essence, we are attempting to determine whether there are samples from a different population present. To address this objective, use of advanced robust influence function-based methods, including PROP, Huber, Tukey's bisquare and MCD method were suggested. These and other robust outlier tests will be available in SCOUT in mid-November, 2008. More traditional outlier tests are not recommended (e.g., Dixon, Rosner) because they are themselves influenced by the presence of the same outliers the tests are attempting to detect (suffer from masking effects).

4. If the data points appear to be significant outliers, then calculate the UTL both with and without these data points, to determine whether they make a significant difference. It should also be noted that robust and resistant estimation methods (also in Scout 2008) are available that automatically compute various statistics of interest (e.g., UCLs, UPLs, UTLs) in which the influence of outliers on the statistics of interest has been reduced to a negligible level.

5. If the outliers affect the result, the workgroup will make a policy call on whether to include them.

- Outliers for multivariate data (sums of compounds) can be identified using a reference envelope approach – basically a threshold in p-space – or by collapsing the data to a single variable such as MDs and using the above approach.

<u>Question 9</u>. If the data can be fit to more than one distribution, what can be used to assist in the selection of the most appropriate distribution? (For instance, Singh and Singh recommend a distribution for calculating a UCL, but do not do so for other statistics such as upper tolerance limits.)

- ⇨ Use the normal distribution whenever possible and appropriate
- ⇨ Do not use transformations to achieve normality, as decisions should be made in the original space
- ⇨ Distributional assumptions may not matter much and it is preferable to use non-parametric methods in most cases

Consensus Points:

- The above three points were agreed upon. Due to the increased availability of nonparametric methods, there is not much need to work with distributions if they aren't normal to begin with.

Question 10. Many project decisions involve iterative removal of contaminated areas or volumes until the remaining sediments are within background. For example, in a dredging project, individual DMMUs may be removed until the remaining ones are clean enough to fall within background (see provided data set as an example). For a cleanup project, areas may be cleaned up through capping or removal until the remaining areas fall within background. Is there any guidance you can provide on quickly identifying the portion of a project distribution that is likely to fall within the background distribution (or alternatively, that which is not)?

- ⇨ Individual sample comparisons to a threshold eliminates this issue
- ⇨ For distribution comparisons, no easy approach, likely an iterative process (removing samples one at a time or in groups and recalculating whether the populations are the same)

Consensus Points:

- The experts agreed on the above two points.

- In addition, the panelists strongly recommended graphical methods to identify data that may not fall within the background distribution, such as multiple Q-Q plots, side-by-side box plots, and comparison of cumulative distribution functions (to identify cases where the majority of the distribution may be the same but the tails may differ). ProUCL (2007) and SCOUT (2008) have several graphical methods that can be used for comparisons of two or more groups of data sets.

Multi-Step/Multi-Part Comparisons

- The panelists did not recommend multi-step or multi-part comparisons, such as a comparison to the mean as well as a comparison to an upper threshold. It was expressed that this type of multiple test reflects uncertainty as to the goal of the comparison and leads to confusion in the results. A single test, properly selected and designed, should meet the comparison needs.

- The exception is where a comparison of population distributions is conducted, and shows that the distributions are different. A subsequent evaluation may be needed to identify the specific samples resulting in that difference, in order to remove them. A threshold that has previously been computed could be used, or a graphical method as discussed above along with iterative population comparisons.

<u>Question 11</u>. Given the recommendations above, what statistical programs are available to make these comparisons, and what are their pros and cons, taking into account the following considerations:

- Available to both project proponents and regulatory parties. The same (or very similar) software should be used by the regulated as well as the regulatory agencies to allow project proponents to propose projects that are likely to be acceptable, and so that mutual checking can occur.
- Cost
- Need for front ends, batching routines, or other modifications for general use
- Support and updates
- Easy to use, or training readily available

S+

- ⇨ Commercial, expensive ($2000+)
- ⇨ Very powerful, customizable
- ⇨ User interface

SPSS

- ⇨ Commercial, expensive ($2000+)
- ⇨ Powerful, more traditional statistics
- ⇨ Not customizable
- ⇨ User interface

R

- ⇨ Free
- ⇨ Less easy for non-expert use
- ⇨ Powerful, customizable

ProUCL

- ⇨ Free
- ⇨ Distributed by EPA
- ⇨ Detailed user's manual and technical guide
- ⇨ User interface
- ⇨ Handles nearly all the questions raised

Scout

- ⇨ Free
- ⇨ Distributed by EPA
- ⇨ Includes ProUCL among many other modules (e.g., additional graphical tools, outlier analyses, PCA, multivariate approaches…)
- ⇨ Coming in mid-November

- Of these, Scout and R were recommended most highly as free programs – Scout for the more casual user and R for the more expert user. Until Scout comes out, ProUCL has many of the functions discussed at the workshop.

ATTACHMENT A

Information for Statistical Experts' Workshop
Regional Sediment Evaluation Team

A. Background

1. In regional dredging programs, a dredged material management unit (DMMU [1]) has traditionally been compared with a background (or reference) data set or sample to make decisions regarding material suitability for open-water or confined disposal. For instance, if a sediment is greater than a screening level, toxicity testing is done with the intent to determine whether toxicity for a test sample is statistically significantly different from, and greater than, the toxicity in a reference sample by a margin (e.g.) 20%. For bioaccumulative compounds, if sediment exceeds a threshold concentration, then bioaccumulation testing has been done for comparison to risk-based threshold values in tissues. An example of the hypothesis formation is shown in the attached PSDDA clarification paper.

2. However, the treatment of bioaccumulative compounds is changing due to several recent developments. One is that some risk-based threshold values for sediments fall below background concentrations for some compounds. Another is that bioaccumulation testing in both the dredging and cleanup programs is increasingly needed to confirm bioavailability of some compounds in sediment. For both cases, comparison to a background data set may be desirable. However, the regulatory process often depends upon "bright lines" to facilitate decision-making. Should the comparison not be able to generate a "bright-line" or "reason-to-believe" screen, one outcome would be less regulatory predictability for the project proponent and greater staff time for the regulators to review projects. However, we recognize that this may not be the most statistically robust approach. In addition, cleanup sites may have more flexibility to use alternative approaches, such as comparisons of distributions.

3. For this workshop, assume an existing background data set for comparison (such as sediment or tissue) independent of a proposed dredging or cleanup project. The background data set for various chemicals/areas may vary in size and completeness, and will also vary greatly in the percentage of non-detects across the data set. Finally, data sets will often include classes of chemicals with joint modes of toxic action, such as carcinogenic dioxin-like compounds (chlorinated dioxins and furans, polychlorinated biphenyls); other carcinogenic compounds (e.g., some PAHs); and narcotic compounds (again, PAHs). For these classes of compounds, there is additivity of different chemicals, e.g., dioxin congeners after multiplying by a "toxicity equivalence factor." An example data set consisting of dioxins is included separately to illustrate the type of data that could be present, with some of the potential issues.

4. Also, assume that a dredging project data set could contain 1 to 3 samples per DMMU, and that multiple DMMUs comprise the project. Most dredging projects have between 3 and 10

[1] This is a decision unit consisting of 100s to 1000s of cubic yards of material being tested for suitability for open-water (or other) disposal.

analytical samples, total. For cleanup sites, the data set may consist of a larger data set (10-50 samples, for example).

5. The desired outcome from this workshop is to renew and refine our understanding of the statistical approach for comparing project sediments to a background data set in order to efficiently and defensibly support regulatory decisions. The following questions are suggested to guide the statistical experts

B. Questions for the Experts

The overall question is how to perform a comparison of project sediments (either sample-by-sample or as a distribution) to a background distribution or threshold to determine whether the project samples exceed the background population.

Single-sample vs. Population Comparisons

Question 1. What are the statistical advantages or disadvantages of using a background distribution vs. a "bright-line" threshold based on the background distribution for comparison to project samples? Does this answer depend on the size of the background data set and/or its distribution? If so, how?

Question 2. What are the statistical advantages or disadvantages of comparison of single test samples vs. test sample populations to background? Does this answer depend on the size of the project data set and/or its distribution? If so, how?

Question 3. If the agencies chose to use a bright-line threshold for the background distribution (UCL, UPL, UTL, upper percentile, etc.), what alternatives are available, and what are their pros and cons? Does the recommended threshold depend on the size of the background data set and/or its distribution? How do the possible thresholds compare in terms of minimizing false negatives vs. false positives? An example used in San Francisco is attached for your review.

Distributions of the Background and Project Data Sets

Question 4. If the data can be fit to more than one distribution, what can be used to assist in the selection of the most appropriate distribution? (For instance, Singh and Singh recommend a distribution for calculating a UCL, but do not do so for other statistics such as upper tolerance limits.)

Question 5. If the agencies chose to use a population-based comparison, how does the type of distribution(s) affect the selection of an exact test to compare the project data with the background data? Which parametric or nonparametric tests should be used for various distributions? Is a flow diagram needed to guide project proponents and staff in making these decisions? If so, what would that look like?

Question 6. What is the role of outlier tests, particularly in establishing a background data set, and which method would be most appropriate?

Multi-step Comparisons

Question 7. Is there value to conducting a multi-step comparison; for example, a bright-line threshold above which individual project samples are excluded, followed by a comparison of the remaining sample distribution to the background distribution?

Question 8. Is there value to conducting a multi-part comparison; for example, a population-based comparison to the mean as well as a threshold that individual samples cannot exceed?

Question 9. Many project decisions involve iterative removal of contaminated areas or volumes until the remaining sediments are within background. For example, in a dredging project, individual DMMUs may be removed until the remaining ones are clean enough to fall within background (see provided data set as an example). For a cleanup project, areas may be cleaned up through capping or removal until the remaining areas fall within background. Is there any guidance you can provide on quickly identifying the portion of a project distribution that is likely to fall within the background distribution (or alternatively, that which is not)?

Treatment of Non-detects

Question 10. What alternatives are available for addressing non-detects (in either distribution), what are their pros and cons, and which would you recommend for individual compounds?

- Does the answer depend on the percentage of non-detects in the data set, and if so, how?
- What is the effect upon the robustness of the testing of replacing non-detected values with reporting limits, ½ reporting limits, 0, or a Regression-on-Order (ROS) statistic while performing a parametric test?

Question 11. Using the data set provided as an example, how should non-detects within summed classes of compounds be addressed? For instance:

- Congeners may be detected at various frequencies, from 0-100%. However, because they are all typically added together on a station-by-station basis to calculate an overall TEQ, a single type of non-detect substitution is typically employed (e.g., ½ the reporting limit). Is this appropriate? If so, what is the best substitution to use, and if not, what alternative approaches are available?

Question 12. How does the percentage of non-detects in the data set affect the selection of a statistical test? Should this be factored into the flowchart?

Hypothesis Testing

Question 13. Given the answers to the questions above, what specific form of the hypothesis would be recommended, including a) null hypothesis, b) alternate hypothesis, and c) α and β levels. (This will be developed at the workshop, based on the discussion to that point.)

Recommended Statistical Software

Question 14. Given the recommendations above, what statistical programs are available to make these comparisons, and what are their pros and cons, taking into account the following considerations:

- Available to both project proponents and regulatory parties. The same (or very similar) software should be used by the regulated as well as the regulatory agencies to allow project proponents to propose projects that are likely to be acceptable, and so that mutual checking can occur.
- Cost
- Need for front ends, batching routines, or other modifications for general use
- Support and updates
- Easy to use, or training readily available

ATTACHMENT B
Attendance List

Teresa Michelsen, Facilitator[1,2]
Loveday Conquest, Expert Panelist, University of Washington
Dennis Helsel, Expert Panelist, Practical Stats
Lorraine Read, Expert Panelist, TerraStat Consulting Group
Anita Singh, Expert Panelist, Lockheed Martin

Shannon Ashurst, ENSR
Jeremy Buck, US F&W[1]
Merv Coover, ENSR
Nancy Harney, EPA Region 10
Erika Hoffman, EPA Region 10[1,2]
Laura Inouye, WA Dept. of Ecology[1,2]
John Malek, Parametrix[1]
Roger McGinnis, Hart Crowser
Lucas Menendez, TestAmerica
Mandy Michalsen, Seattle District COE
Nancy Musgrove, Windward
Gene Revelas, Integral Consulting
Paul Seidel, Oregon DEQ[1]
Alice Shelly, TerraStat
Stephanie Stirling, Seattle District COE[1]
Lucinda Tear, Windward
Todd Thornburg, Anchor[1]

[1] RSET Bioaccumulation Subcommittee member
[2] DMMP Dioxin Workgroup member