February 12, 2009

Evaluation of Bold and Main Basin Plus Reference Area Dioxin Data

1. Purpose.

This memo describes statistical considerations of the two data sets of polychlorinated dioxin and furan data for the DMMP agencies' deliberation in the process of selection of an appropriate background threshold value (BTV) as a suitability criterion for open-water disposal of dredged material in Puget Sound. The memo accompanies spreadsheets with all data used to illustrate the calculations. These are spreadsheets in Microsoft ™ Excel and the freely-available EPA software, ProUCL v. 4 [1].

The "Bold data set" consists of 70 stations, and 5 had laboratory duplicates; but only the primary samples are considered here. The Bold data include a number of samples that are from reference bays sampled as part of that investigation. Also, 27 other data points deemed to be comparable to the Bold reference bays have been included, and this larger data set is named the "Bold plus existing main basin and reference bay data." Both data sets are evaluated here, for different purposes. The Bold dataset is being described here because it is new and has been recently posted on the DMMP website, as a part of the investigation findings. The larger data set is intended for use in the BTV deliberation.

2. Methods and Materials.

2.1 The first step is calculation of nonparametric Kaplan-Meier (K-M) sample TEQ sums for all the below-detection-limit (BDL). This is done in accordance with the procedure suggested by Dr. Dennis Helsel at Practicalstats.com (formerly with USGS) during the Regional Sediment Evaluation Team Statistics Workshop in November, 2008 (synopsis attached as Appendix A). This procedure, which is currently awaiting 2009 publication by Dr. Helsel in the *Journal of Environmental Science and Technology,* was selected due to the presence of many samples with BDL observations. EPA's ProUCL software is used in conjunction with Crystal Ball ™ and Excel ™ for organizing and summarizing the data.

Appendix B documents the manipulations necessary to accomplish the K-M sum estimation and background calculations. It also points out the portions of the spreadsheets that contain the manipulations and calculations.

2.2 The second step is statistical summary of K-M TEQ sum results for the datasets. This step includes graphic comparison of the K-M TEQ sample sums to the sample sums derived from the ND=0 and ND=0.5DL approaches. Statistical experts at the cited workshop recommended against these two latter substitutions; however, even if these substituted statistics are not to be used for the statistics, they are the means by which dioxin/furan data are commonly

[1] http://www.epa.gov/nerlesd1/tsc/software.htm

summarized, and so are provided for comparison. The summary below includes the parametric estimation of the median and the 25th, 80th and 90th percentiles (coverages) of the distribution of the data. ProUCL does not calculate all of these statistics, so Excel statistical functions have been used for some, following recommendations for distributions derived from ProUCL and Crystal Ball ™.

2.3 The third step is calculation of example BTVs for the Bold and Bold plus existing main basin and reference bay datasets for agency deliberation. This is done in 2 phases.

2.3.1 Step 3 phase 1 is the nonparametric estimation of the 90th and 95th upper confidence intervals of the 80th , 85th , and 90th percentile coverages. Nonparametric methods such as Kaplan-Meier were recommended in the RSET statistics workshop. These were calculated using ProUCL.

2.3.2 Step 3 phase 2 is comparison of the nonparametric statistics to the parametric 90th percentile suggested in MTCA "Statistical Guidance or Site Managers." This step was intended to provide information for the agencies to assist in selecting an appropriate nonparametric statistic.

3. Results of Step 1, calculation of K-M sum of TEQ for individual samples.

3.1 Table 1 displays results of the three methods of determining sample TEQ sums.

3.2 Figure 1 shows a quantile-quantile (Q-Q) plot of these results. The K-M estimate falls between the ND=0 and ND=0.5DL estimates. The K-M method generally returns somewhat lower summed TEQs when the magnitude of the detected congeners is at or below about 2 ng TEQ/kg (Figure 2). In this figure, the y axis is the difference between the ND=0.5DL and K-M results, divided by the ND=0.5DL result; thus, it could be read as “fractional difference between estimates.” Above 2 or 2.5 ng TEQ/kg, there is less difference between the estimates. (This may illustrate limitations of interpretation below these values.)

Table 1. Results of 3 Methods of Estimating Sample TEQ Sums. (Yellow cells are Bold data; blue cells are other reference areas.)

Station	K-M TEQ	K-M Note	ND=0.5DL	ND=0
AI_1	0.423		0.564	0.358
AI_11_C	0.044		0.258	0.023
AI_13_C	0.452		0.568	0.376
AI_20_C_GS	0.619		0.670	0.567
AI_5_C	0.408		0.544	0.329
AKB02-2005	1.625		1.685	1.564
AKB02-2008	1.195		1.483	0.954
AKB03-2005	2.618		2.637	2.560
AKB03-2008	1.610		1.887	1.398
AKP01-2005	6.794	a	6.794	6.794
AKP02-2005	2.346		2.420	2.266
AKP02-2008	2.431		2.466	2.328
AKP03-2005	4.404	a	4.404	4.404
AKP03-2008	2.091		1.956	1.377
AKP04-2005	4.403		4.474	4.276
AKP04-2008	1.887		2.271	1.796
AKT01-2005	2.890		2.972	2.794
AKT01-2008	1.503		1.793	1.297
CPS_0	1.853		1.952	1.741
CPS_1	2.193		2.217	2.127
CPS_3	1.331		1.366	1.275
CPS_4	0.949		1.036	0.844
CPS_5	0.655		0.733	0.582
CR02	2.278		2.290	2.285
CR23	0.728		0.743	0.701
CR23W	0.440		0.479	0.333
CR24	0.780		0.784	0.770
Grays OM90 Sequim	2.890		2.026	0.030
Grays OM90 West Beach	0.149		4.065	2.655
HC_0	0.886		1.015	0.795
HC_1	0.802		0.871	0.721
HC_2	0.774		1.154	0.584
HC_3	0.444		0.646	0.355
HC_6	0.493		0.769	0.359
MSMP-43	0.180		0.274	0.138
NCPS_0	0.646		0.758	0.532
NCPS_1	0.079		0.261	0.014
NCPS_2	1.068		1.231	0.923
NCPS_3	0.675		0.763	0.617
NCPS_4	0.298		0.498	0.243
PASED08 RF03A	0.857		0.857	0.856
PASED08-RF01A	0.041		0.062	0.034
PASED08-RF02A	0.055		0.101	0.041
PSPS_1	2.040		2.194	1.879
PSPS_2	2.686		2.980	2.387
PSPS_3	0.862		1.056	0.749
PSPS_8	0.105		0.327	0.067
PSPS_9	1.464		1.670	1.325
R_CAR_0	0.598		0.823	0.444
R_CAR_1	1.040		1.215	0.865
R_CAR_4	0.838		1.068	0.652
R_CAR_5	5.066		5.152	4.920
R_CAR_6_C	0.211		0.510	0.135
R_DAB_0	0.257		0.581	0.151
R_DAB_1	1.578		1.685	1.486

R_DAB_2	1.443		1.789	1.229
R_DAB_5	1.532		1.848	1.297
R_DAB_7_C	1.202		1.225	1.156
R_HOL_0	0.121		0.295	0.080
R_HOL_1	0.372		0.401	0.332
R_HOL_3	0.100		0.241	0.061
R_HOL_4	1.199		1.434	1.020
R_HOL_7	0.862		0.946	0.778
R_SAM_0	1.321		1.390	1.261
R_SAM_1	1.561		1.604	1.523
R_SAM_3	1.324		1.381	1.285
R_SAM_4	0.877		1.035	0.814
R_SAM_5	1.836		1.888	1.753
RAYONR05S1-01	0.051		0.155	0.027
RAYONR05S1-02	0.110		0.181	0.090
RAYONR05S1-03	0.155		0.365	0.095
SamishRef	2.439	a	2.439	2.439
SCPS_1	3.349		3.387	3.261
SCPS_10_C	1.088		1.158	1.017
SCPS_2	0.508		0.566	0.442
SCPS_3	0.177		0.340	0.126
SCPS_5	3.655		3.737	3.471
SJF_10_C	0.323		0.465	0.247

SJF_12_C_GS	1.678		1.747	1.575
SJF_2	0.275		0.500	0.212
SJF_3	0.163		0.420	0.123
SJF_9_C	0.536		0.803	0.417
SJI_0	0.677		0.991	0.516
SJI_1	0.828		0.899	0.770
SJI_20_C_GS	1.149		1.341	1.019
SJI_3	0.445		0.667	0.352
SJI_8_C	0.556		0.722	0.439
SPSB_0	1.457		1.577	1.403
SPSB_1	1.265		1.384	1.164
SPSB_2	2.142		2.271	1.968
SPSB_3	0.192		0.421	0.132
SPSB_8_C	0.073		0.319	0.029
SS_0	8.347		8.311	8.121
SS_1	0.393		0.525	0.331
SS_2	1.331		1.367	1.258
SS_8_C	1.103		1.186	0.954
SS_9_C	11.594		11.626	11.354

[a] These samples had no congeners BDL. There was no corresponding K-M value, and the standard sum of TEQs is shown.

Figure 1. Q-Q Plot of Three Methods for Calculating Dioxin TEQ

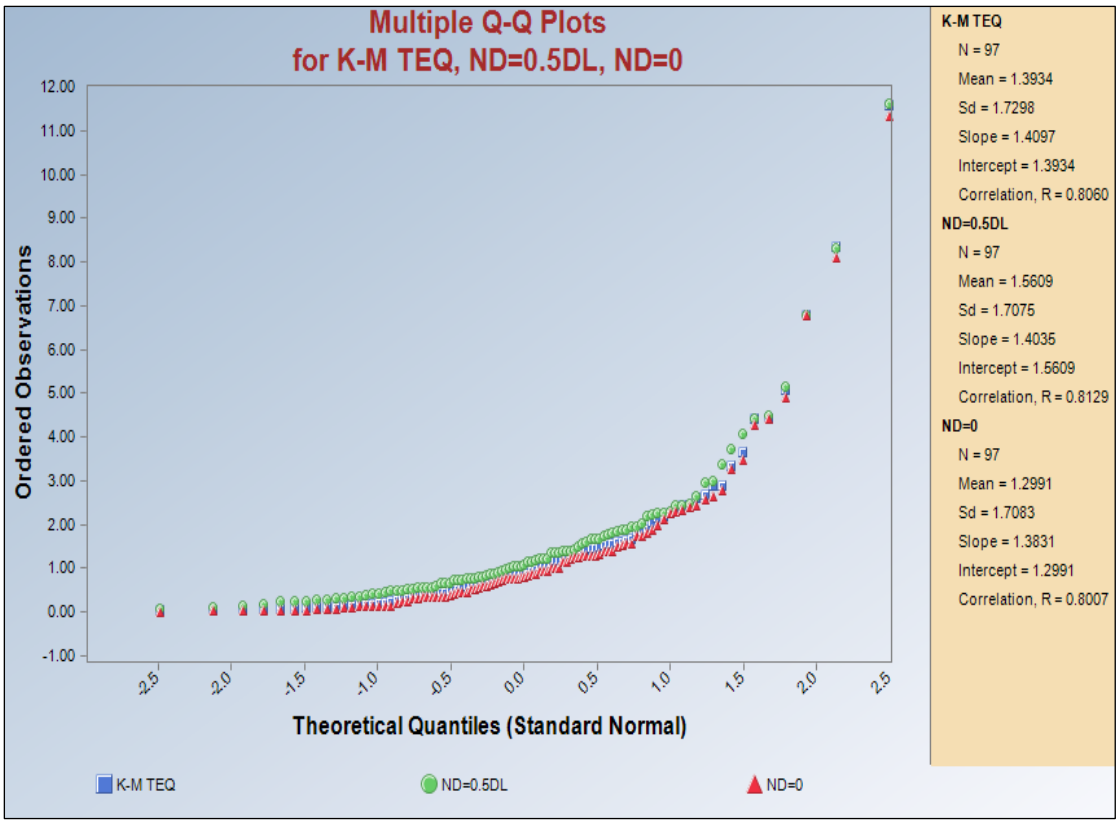
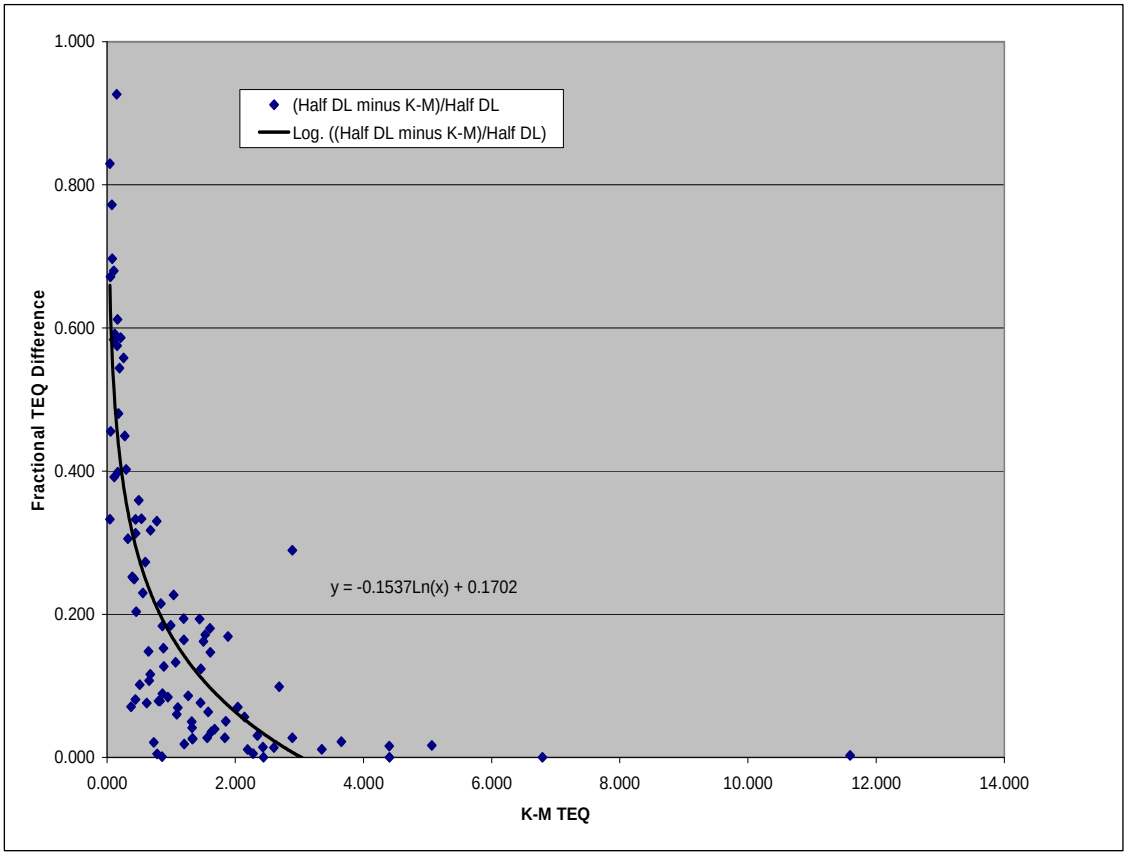


Figure 2. Relative Differences Between K-M and ND=0.5DL Methods of Estimation of TEQ Sum



4. Results of Step 2.

Table 2. Summary Statistics of Kaplan-Meier TEQ Sums (see spreadsheet "Crystal Ball Stats Estimates.XLS")

	Bold Only	**Note**	Bold Plus Existing Main Basin and Reference Bays	**Note**
25th Percentile	0.36	a	0.42	b
Median	0.85		0.88	
80th Percentile	1.82	a	2.26	b
85th Percentile	2.21	a	2.68	b
90th Percentile	2.86	a	3.24	b

(a) - Lognormal recommended by ProUCL. Crystal Ball also recommends this, and fits it to a mean of 1.26 and SD of 1.66. These values are derived from the cumulative fitted distribution in Crystal Ball.
(b) – Gamma distribution recommended by ProUCL;but exponential recommended by Crystal Ball, and the latter was used, with a rate of 0.72. Latter chosen for the statistical summary.

5. Results of Step 3.

5.1 Development of K-M Nonparametric Estimates as Potential BTVs.

Table 3. Nonparametric estimation of the 90th and 95th upper confidence intervals for 80th to 90th percentile coverages

Coverage	**Confidence**	**Bold Only**	**Bold Plus Existing Main Basin and Reference Bays**
80th Percentile	90% Confidence Interval	1.84	2.28
	95% Confidence Interval	1.85	2.35
85th Percentile	90% Confidence Interval	2.14	2.62
	95% Confidence Interval	2.19	2.89
90th Percentile	90% Confidence Interval	3.35	3.66
	95% Confidence Interval	3.35	3.66

5.2 Comparison of Bold Plus Existing Main Basin and Reference Bays statistics from Tables 2 and 3.

Using Crystal Ball, it was possible to relate Table 3 values to the precise cumulative percentile of the distributions in Table 2.

- The Table 3 K-M value of 2.60 ng TEQ/kg representing the 85th percentile/90th confidence interval corresponds to the 85th percentile of the exponential distribution fitted by Crystal Ball.
- The Table 3 K-M value of 2.89 ng TEQ/kg representing the 85th percentile/95th confidence interval corresponds to the 86th percentile of the exponential distribution fitted by Crystal Ball.
- The Table 3 values of 3.66 ng TEQ/kg correspond to the 93rd percentile of the exponential distribution fitted by Crystal Ball.

5.2 Duplicate Sample Variability from the Bold Main Basin and Reference Bays data set.

It is appropriate to consider the influence of analytical variability on the result. In these data sets, there are six lab duplicates. [2]

Table 4. Duplicate Results from Bold and Anderson-Ketron 2005 Dioxin Samples.

	K-M Sum	**Relative Percent Difference** [3]	**Difference (K-M Sums)**
	ng TEQ/kg	%	ng TEQ/kg
CPS_3_Dup	1.55		
CPS_3	1.33	5%	0.22
HC_2_Dup	3.33		
HC_2	0.77	1%	2.56
NCPS_2_Dup	0.92		
NCPS_2	1.07	15%	0.14
PSPS_1_Dup	0.95		
PSPS_1	2.04	17%	1.09
SPSB_0_Dup	1.57		
SPSB_0	1.46	5%	0.12
AKB03-2005	2.601		
AKB03-2005 Dup	2.159	1%	0.44
Range		1-17%	0.12-2.56
Geometric mean		5%	0.41

These results reflect both the analytical uncertainty and the sampling heterogeneity associated with the reported values.

[2] An additional duplicate pair , AKP01, was collected from the 2008 Anderson-Ketron data setbut was not used in this evaluation as this station was not regarded as un-impacted by recent dredging events. However, for information, the difference of K-M sums was 0.71 ng TEQ/kg, and is within the range of the values shown in the right-hand column.

[3] The absolute difference of the 2 values divided by their average, expressed as a percent.

Appendix A. Statistics Workshop Record

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.

<u>Question 3.</u> 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.

Question 4. 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.

<u>Question 6</u>. 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. Appropriate software (e.g., Scout 2008) can then be used to calculate project Mahalanobis distances (MDs) from the centroid of the background data set in p-dimensional space. 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.

Question 7. 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) ≤ 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).

<u>Question 8</u>. 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.

Question 9. 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.

Question 11. 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.

APPENDIX A - ATTACHMENT 1

Information for Statistical Experts' Workshop
Regional Sediment Evaluation Team

A. Background

1. In regional dredging programs, a dredged material management unit (DMMU [4]) 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] 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.

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 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

APPENDIX A - 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

Appendix B. Instructions for Using Pro-UCL to Create K-M Sums for Samples, and Key to the Spreadsheets.

Appendix B. Instructions for Using Pro-UCL to Create K-M Sums for Samples, and Key to the Spreadsheets.

1. Data as received resembles this. The order of the congeners is unimportant, but the pairing with the TEF (as well as use of the WHO 2005 TEFs) is critical.

Sample ID		**AI_1**		**AI_11_C**	
Analyte (pg/g)	**TEF**	**7/31/2008**	**Q**	**8/1/2008**	**Q**
% Solids		68.6		75.1	
2,3,7,8-TCDD	1	0.144	U	0.141	U
1,2,3,7,8-PeCDD	1	0.21	U	0.144	U
1,2,3,4,7,8-HxCDD	0.1	0.227	U	0.209	U
1,2,3,6,7,8-HxCDD	0.1	0.431	J	0.217	U
1,2,3,7,8,9-HxCDD	0.1	0.349	J	0.225	U
1,2,3,4,6,7,8-HpCDD	0.01	4.19	J	0.367	J
OCDD	0.0003	24.5		2.22	U
2,3,7,8-TCDF	0.1	0.404	J	0.192	J
1,2,3,7,8-PeCDF	0.03	0.122	U	0.118	U
2,3,4,7,8-PeCDF	0.3	0.269	J	0.117	U
1,2,3,4,7,8-HxCDF	0.1	0.264	J	0.174	U
1,2,3,6,7,8-HxCDF	0.1	0.317	J	0.163	U
1,2,3,7,8,9-HxCDF	0.1	0.275	U	0.243	U
2,3,4,6,7,8-HxCDF	0.1	0.315	J	0.171	U
1,2,3,4,6,7,8-HpCDF	0.01	1.92	J	0.225	U
1,2,3,4,7,8,9-HpCDF	0.01	0.383	U	0.348	U
OCDF	0.0003	2.57	J	0.54	J

2. Preparation for ProUCL.

2.1 Copy the data to a new worksheet. I called mine "UCL Prep 1." In the above example, no TEQs have been calculated. If they had been, and if the formulas used a substitution such as ND=0.5DL, those columns should be eliminated now, as should unnecessary data (such as the % solids row).

2.2 A new column should be inserted to the right of the raw concentration data. Then, the sample identifier (AI_1) should be copied over the new column and the Q column for that sample. It will look like this.

	A	B	C	D	E
1	**Sample ID**		**AI_1**		
2	**Analyte (pg/g)**	**TEF**	**7/31/2008**	**AI_1**	**AI_1**
3	2,3,7,8-TCDD	1	0.144		
4	1,2,3,7,8-PeCDD	1	0.21		U
5	1,2,3,4,7,8-HxCDD	0.1	0.227		U
6	1,2,3,6,7,8-HxCDD	0.1	0.431		U
7	1,2,3,7,8,9-HxCDD	0.1	0.349		J
8	1,2,3,4,6,7,8-HpCDD	0.01	4.19		J
9	OCDD	0.0003	24.5		J
10	2,3,7,8-TCDF	0.1	0.404		
11	1,2,3,7,8-PeCDF	0.03	0.122		J
12	2,3,4,7,8-PeCDF	0.3	0.269		U
13	1,2,3,4,7,8-HxCDF	0.1	0.264		J
14	1,2,3,6,7,8-HxCDF	0.1	0.317		J
15	1,2,3,7,8,9-HxCDF	0.1	0.275		J
16	2,3,4,6,7,8-HxCDF	0.1	0.315		U
17	1,2,3,4,6,7,8-HpCDF	0.01	1.92		J
18	1,2,3,4,7,8,9-HpCDF	0.01	0.383		J
19	OCDF	0.0003	2.57		U

2.3 Next, create a formula to multiply the concentration times the TEF. In the present case, in cell D3 the formula is =C3*B3. The $ code is an absolute reference, and permits you to copy this to other cells for other samples at the same row. Copy this formula down that column. Then, you need to fix it so that it always references the correct TEF to the left of the concentration. E.g., for cell D13, the formula should be =C13*B13. Now it resemble this:

Sample ID		**AI_1**		
Analyte (pg/g)	**TEF**	**7/31/2008**	**AI_1**	**AI_1**
2,3,7,8-TCDD	1	0.144	0.144	
1,2,3,7,8-PeCDD	1	0.21	0.21	U
1,2,3,4,7,8-HxCDD	0.1	0.227	0.0227	U
1,2,3,6,7,8-HxCDD	0.1	0.431	0.0431	U
1,2,3,7,8,9-HxCDD	0.1	0.349	0.0349	J
1,2,3,4,6,7,8-HpCDD	0.01	4.19	0.0419	J
OCDD	0.0003	24.5	0.00735	J
2,3,7,8-TCDF	0.1	0.404	0.0404	
1,2,3,7,8-PeCDF	0.03	0.122	0.00366	J
2,3,4,7,8-PeCDF	0.3	0.269	0.0807	U
1,2,3,4,7,8-HxCDF	0.1	0.264	0.0264	J

1,2,3,6,7,8-HxCDF	0.1	0.317	0.0317	J
1,2,3,7,8,9-HxCDF	0.1	0.275	0.0275	J
2,3,4,6,7,8-HxCDF	0.1	0.315	0.0315	U
1,2,3,4,6,7,8-HpCDF	0.01	1.92	0.0192	J
1,2,3,4,7,8,9-HpCDF	0.01	0.383	0.00383	J
OCDF	0.0003	2.57	0.000771	U

2.4 ProUCL needs to have columns that show the result with ND=DL, and a code that says whether it is detected or not, i.e., 1 for detected and 0 for nondetected. First, look over your data set and identify the codes that represent nondetected congeners. These are typically U and UJ, but there may be others such as UJK. Then, be sure to highlight a range of data that doesn't include your sample ID names. In one instance, my entry Sequim became SeqOim! Substitute for nondetected codes a "0" using ctrl-F search and replace. It is recommended that you do the longer strings first, such as UJ, because Excel will replace *within* a cell and if you do it in another order (say, U first), it will not see the changed value, OJ, as nondetected. Note: where there are no qualifiers, you have to manually insert "1"s. You may alternatively wish to write "=if" functions to do this entire operation. Upon completion, you should have this:

Sample ID		**AI_1**		
Analyte (pg/g)	**TEF**	**7/31/2008**	**AI_1**	**AI_1**
2,3,7,8-TCDD	1	0.144	0.144	1
1,2,3,7,8-PeCDD	1	0.21	0.21	0
1,2,3,4,7,8-HxCDD	0.1	0.227	0.0227	0
1,2,3,6,7,8-HxCDD	0.1	0.431	0.0431	0
1,2,3,7,8,9-HxCDD	0.1	0.349	0.0349	1
1,2,3,4,6,7,8-HpCDD	0.01	4.19	0.0419	1
OCDD	0.0003	24.5	0.00735	1
2,3,7,8-TCDF	0.1	0.404	0.0404	1
1,2,3,7,8-PeCDF	0.03	0.122	0.00366	1
2,3,4,7,8-PeCDF	0.3	0.269	0.0807	0
1,2,3,4,7,8-HxCDF	0.1	0.264	0.0264	1
1,2,3,6,7,8-HxCDF	0.1	0.317	0.0317	1
1,2,3,7,8,9-HxCDF	0.1	0.275	0.0275	0
2,3,4,6,7,8-HxCDF	0.1	0.315	0.0315	1
1,2,3,4,6,7,8-HpCDF	0.01	1.92	0.0192	1
1,2,3,4,7,8,9-HpCDF	0.01	0.383	0.00383	1
OCDF	0.0003	2.57	0.000771	0

2.5 There are 2 data columns (well, 3, but we're going to get rid of the one with the date in it in a while) that say AI_1. ProUCL has to have a "D_" or a "d_" code in front of the detect/nondetect column. Edit the right-most column to say

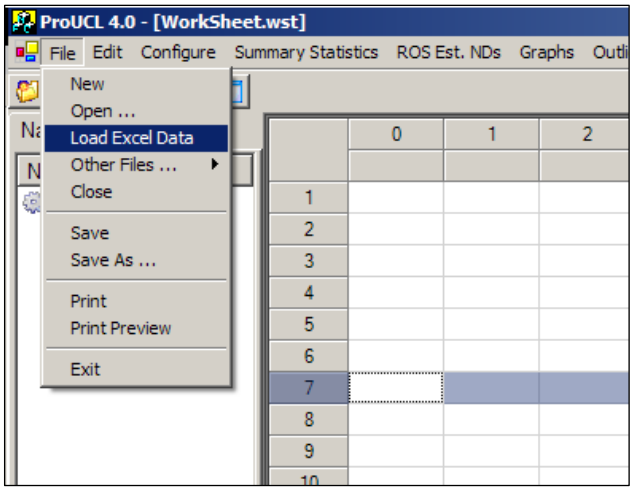
D_AI_1. Repeat this activity across the spreadsheet for every sample. Don't forget at the end of this time to zoom out and assure that there is a detect/nondetect code in every cell. If there isn't, ProUCL will assume the value is "1", and calculate a K-M value for the congener.

2.6 You're ready to copy this to another worksheet. I called mine "ProUCL Prep 2." It is important to copy the data using "Paste Special" and check the box that says "values only." You are going to be delete columns, and you don't want the dependent cells to recalculate. Remove the first three columns and all columns with dates in them. Mine looks like this now for the first two samples. This is almost ready for import to ProUCL.

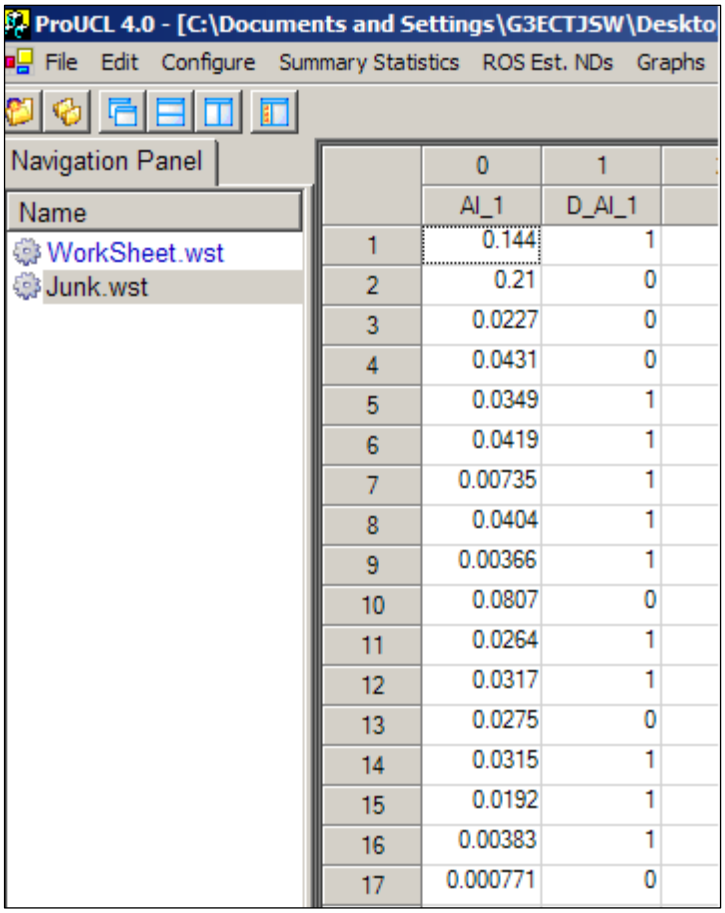
AI_1	D_AI_1
0.144	1
0.21	0
0.0227	0
0.0431	0
0.0349	1
0.0419	1
0.00735	1
0.0404	1
0.00366	1
0.0807	0
0.0264	1
0.0317	1
0.0275	0
0.0315	1
0.0192	1
0.00383	1
0.000771	0

2.7 Copy that worksheet to a new Excel file. ProUCL can import data directly, but it cannot deal with multiple worksheets. Save the Excel file. Since this spreadsheet is just a copy, after you import to ProUCL, you can delete it.

3. Import to ProUCL. Open ProUCL. Here's how you import from your spreadsheet.

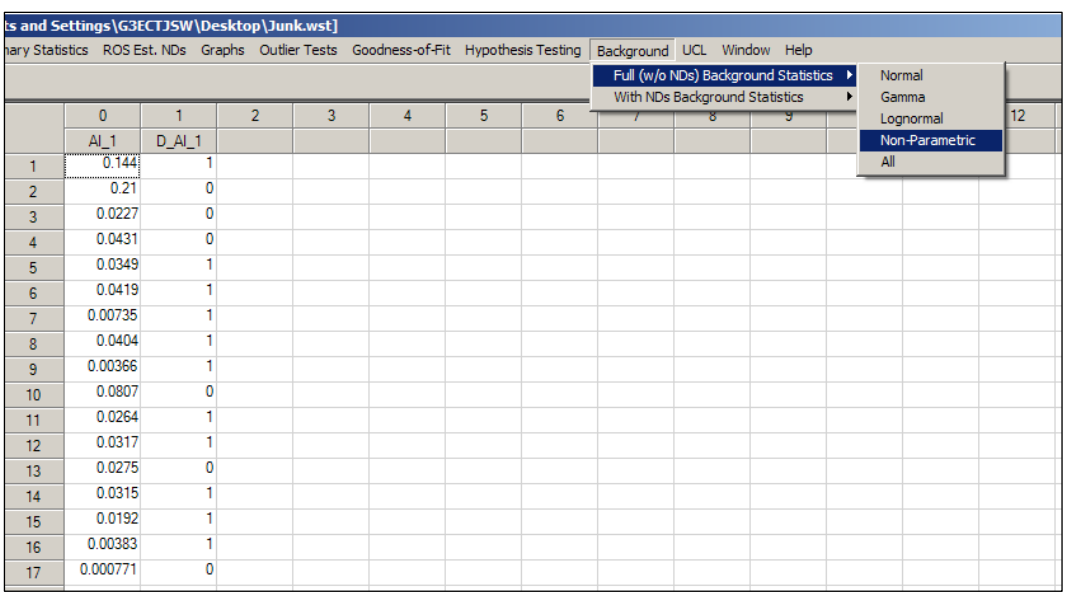


Here is what it should look like for the example data.



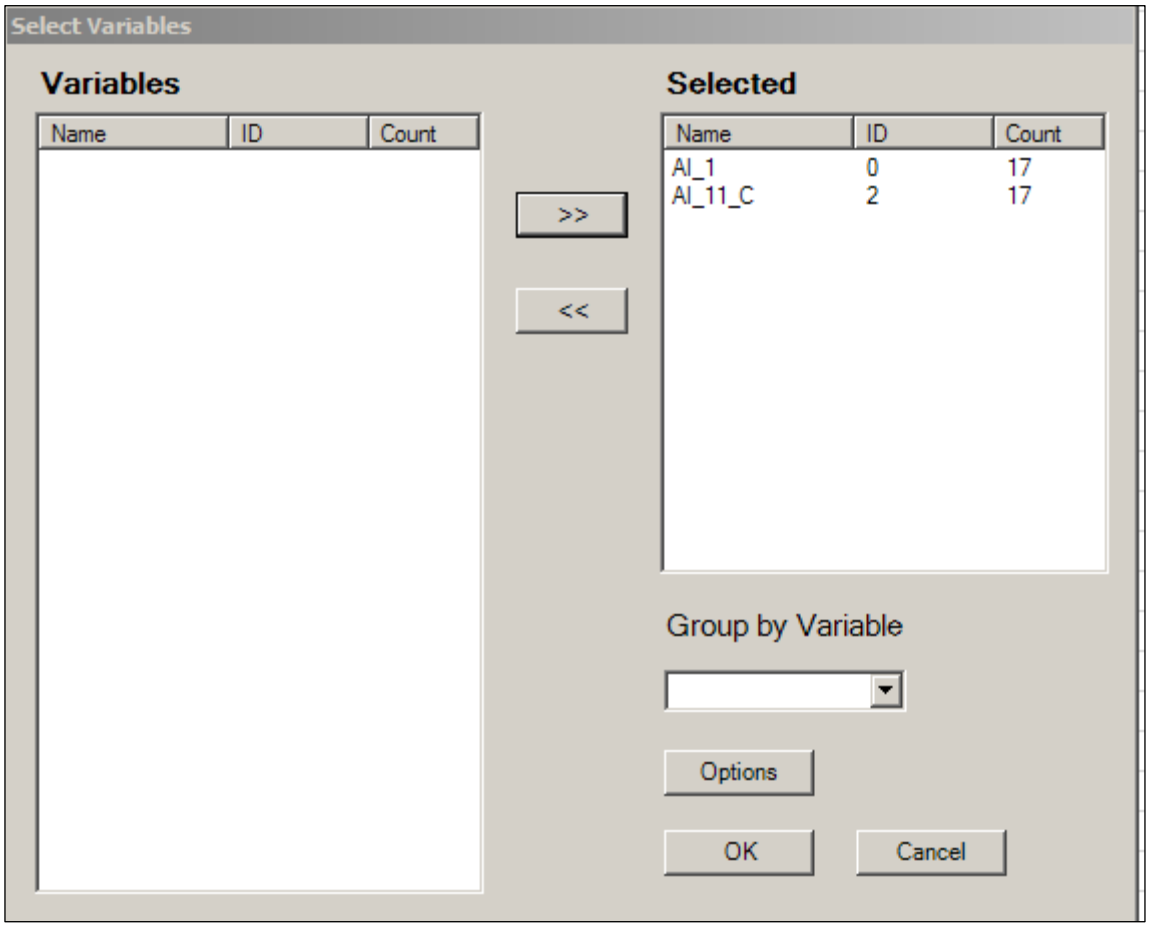
	0	1
	Al_1	D_Al_1
1	0.144	1
2	0.21	0
3	0.0227	0
4	0.0431	0
5	0.0349	1
6	0.0419	1
7	0.00735	1
8	0.0404	1
9	0.00366	1
10	0.0807	0
11	0.0264	1
12	0.0317	1
13	0.0275	0
14	0.0315	1
15	0.0192	1
16	0.00383	1
17	0.000771	0

4. Calculate Sums of TEQs using Kaplan-Meier. You want to create a sum for each sample. ProUCL does not do this directly. Instead it creates a mean for each sample, which when multiplied by the number of congeners (17) is the sum. Mean = sum/n; sum= Mean*n. Here is the process. First, calculate a Background With NDs/Nonparametric:



	0	1
	Al_1	D_Al_1
1	0.144	1
2	0.21	0
3	0.0227	0
4	0.0431	0
5	0.0349	1
6	0.0419	1
7	0.00735	1
8	0.0404	1
9	0.00366	1
10	0.0807	0
11	0.0264	1
12	0.0317	1
13	0.0275	0
14	0.0315	1
15	0.0192	1
16	0.00383	1
17	0.000771	0

ProUCL opens a dialog box showing your two samples. (If you didn't enter the D_AI_1 correctly, that sample will not appear in the Variables window.) Highlight and arrow your samples to the Selected window. Click OK.



The next page shows the output from this example. It's a lot of info. You only care about the K-M mean, highlighted below. (I only show one sample.) I usually just copy this into my Excel spreadsheet of record. (In the present case, it was imported into a worksheet called "ProUCl Outputs No AK," because I did the Anderson-Ketron samples separately.

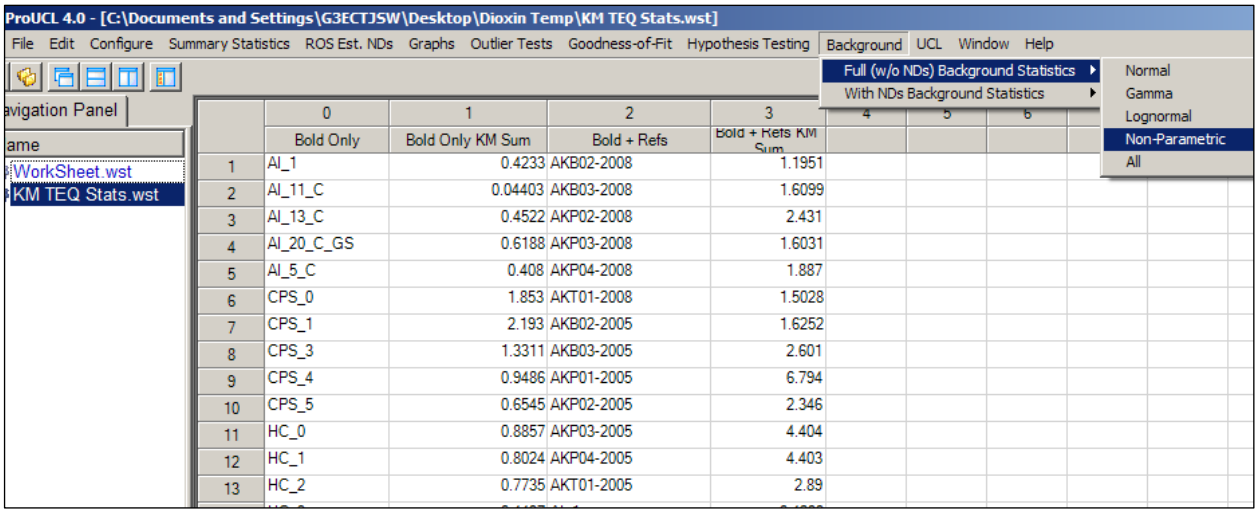
Nonparametric Background Statistics for Data Sets with Non-Detects					
Confidence Coefficient		95%			
Coverage		90%			
Different or Future K Values		1			
AI_1					
Total Number of Data				17	
Number of Non-Detect Data				6	
Number of Detected Data				11	
Minimum Detected				7.71E-04	
Maximum Detected				0.0807	
Percent Non-Detects				35.29%	
Minimum Non-detect				0.00366	
Maximum Non-detect				0.21	
Mean of Detected Data				0.0325	
SD of Detected Data				0.0211	
Mean of Log-Transformed Detected Data				-3.816	
SD of Log-Transformed Detected Data				1.26	
Data appear Normal at 5% Significance Level					
Nonparametric Background Statistics					
95% UTL with 90% Coverage					
Order Statistic				17	
Achieved CC				1	
UTL				0.21	
Warning: Largest Non-detect at Order				17	
95% UPL					
95% UPL				0.21	
Kaplan-Meier (KM) Method					
Mean				0.0249	
SD				0.0216	
Standard Error of Mean				0.00594	
95% UTL 90% Coverage				0.0683	
95% KM Chebyshev UPL				0.122	
95% KM UPL (t)				0.0638	
90% KM Percentile (z)				0.0527	
95% KM Percentile (z)				0.0605	
99% KM Percentile (z)				0.0753	

You will have to copy the K-M mean to your spreadsheet of record (not the one you are going to delete), and multiply it there by 17. The value is the sum of TEQs of the sample.

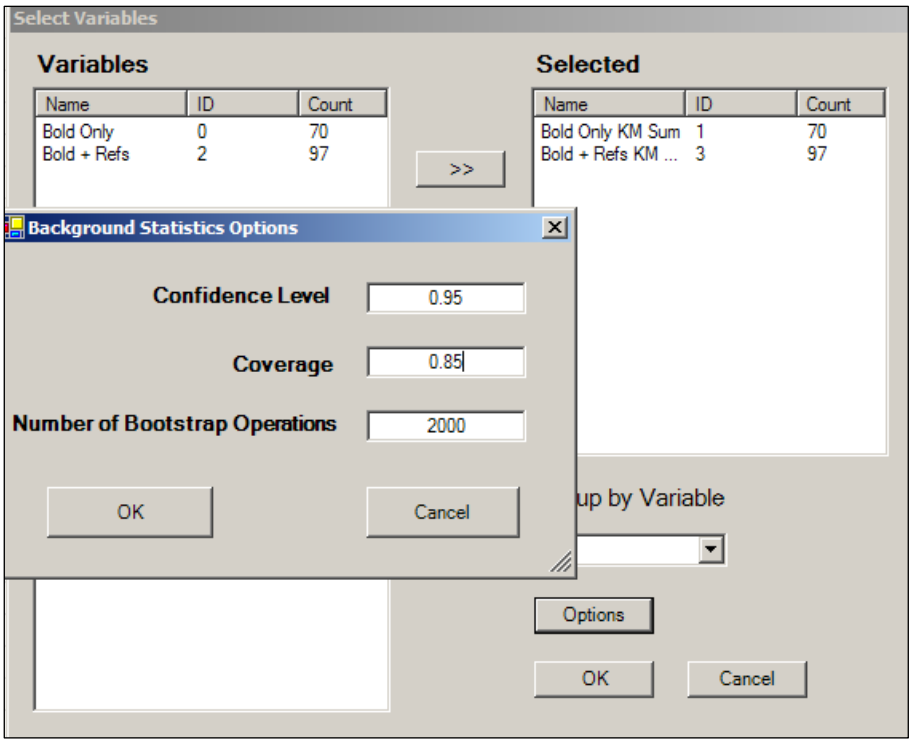
5. Calculate background from a set of sample K-M sums.

5.1 Copy the K-M sums into a blank spreadsheet, and load it as before.

5.2 You will use this setting: calculate a Background/Full (no NDs) Background Statistics/Nonparametric. (You don't need the With NDs form, because you have dealt with that above.)



5.3 As before, select your samples by variable name and arrow them into the Selected box. Supposing you want to calculate a UTL for a 95% confidence interval on the 85th percentile (coverage), the Option box would look like this.



5.4 Your output will look like this. The value you want is highlighted.

Nonparametric Background Statistics for Full Data Sets		
Full Precision	OFF	
Confidence Coefficient	95%	
Coverage	85%	
Number of Bootstrap Operations	2000	
Bold Only KM Sum		
Some Non-Parametric Statistics		
Number of Valid Observations		70
Number of Distinct Observations		68
Minimum		0.044
Maximum		11.59
Second Largest		8.347
Mean		1.268
First Quartile		0.419
Median		0.85
Third Quartile		1.447
SD		1.765
Variance		3.115
Coefficient of Variation		1.392
Skewness		4.103
Mean of Log-Transformed data		-0.305
SD of Log-Transformed data		1.07
Data appear Lognormal at 5% Significance Level		
Non-Parametric Background Statistics		
90% Percentile		2.188
95% Percentile		4.29
99% Percentile		11.59
95% UTL with 85% Coverage		
Order Statistic		64
Achieved CC		0.962
UTL		2.193

6. Key to the Spreadsheets and Worksheets

090210_Bold KM Comparison.xls

This spreadsheet contains all the original data from the Bold, Anderson-Ketron, other Corps' data, and EIM retrievals that formed the basis for the 97-sample data set. Worksheets (from left to right in the spreadsheet) include:

- Summary of Stats Combined Data. This is where Tables 2 and 3 of the main text are documented. This was the last step in the analysis. ProUCL outputs showing the UTLs are shown to the right of the summary table. Key values are blue highlighted.
- K-M Sums. This worksheet summarizes the K-M sample calculations and displays the ND=0.5DL and ND=0 estimates as well. Column H comes from the ProUCL outputs in other worksheets (ProUCL Outputs (A-K) and ProUCL Outputs (No A-K). The reason for 2 worksheets is that the Anderson-Ketron data were worked up before the Bold statistics were.
- Bold data. Self-explanatory.
- Bold data ProUCL Prep 1. This is done in the manner described above (Appendix A, 2.1 ff).
- Bold data ProUCL Prep 2. This is also as described above
- DMMP and Samish Ref Data; DMMP and Samish Ref Data Prep 1; DMMP and Samish Ref Data Prep 2. These are organized as were the Bold data.
- Other EIM refs data; Other EIM Prep 1; Other EIM Prep 2. These are organized as were the Bold data.
- Corps data; Corps Prep 1; Corps Prep 2. These are organized as were the Bold data.
- Anderson-Ketron Data; Anderson-Ketron Prep 1; Anderson-Ketron Prep 2. These are organized as were the Bold data. The last is the load file for ProUCL for the Anderson-Ketron data.
- 2005 and 2008 Anderson-Ketron Data Summary. This has the results of the ProUCL organized with the other methods for summing.
- Combined ProUCL Prep (No A-K). This is the load file for ProUCL for all except the Anderson-Ketron data.
- ProUCL Outputs AK. Self explanatory.
- ProUCL Outputs Except AK. Self explanatory.
- TEFs. This just stores the TEFs in the 2 formats that are found in the data sheets.

Crystal Ball Stats Estimates.XLS

- Bold Only worksheet. This has the Bold data and the fit to the population. (Note that you cannot see cells C1, the assumptions, and C2, the forecast without Crystal Ball. The text describes the assumptions.)
- Bold Plus Ref Bays worksheet. Self explanatory.

ProUCL Worksheets (.WST)

Combined Outputs No AK.WST

- This ProUCL spreadsheet contains the data used in used in doing the K-M sum calculations for all but the Anderson-Ketron data.

AK.WST

- This ProUCL spreadsheet contains the data used in doing the K-M sum calculations for the Anderson-Ketron data.

Bold Stats.WST

- This ProUCL spreadsheet file contains the three estimation methods (K-M, 0.5DL, ND=0) for creation of the Q-Q plot and summary statistics. Note that the columns with "FULL" in them refer to the 97-element combined data set, and BOLD indicate the 70-element data set.