

First memo to D. Smith

Mary



INTEROFFICE CORRESPONDENCE

DATE: September 2, 1993

TO: D. M. Smith, Environmental Engineering and Technology, Bldg. 080, X8636

FROM: M. A. Siders, Geosciences, Bldg. 080, X6933 *M. A.S.*
R. P. Boan, Geosciences, Bldg. 080, X8658 *RB*

SUBJECT: IMPACT OF GILBERT'S METHODS AS SUGGESTED IN HIS JULY 30, 1993 REPORT -
MAS-006-93

We have talked over the impacts of the various tests suggested by Gilbert. In general, we both agree that Gilbert's suggestions are reasonable, and that his "five-phase" process is a good approach. We have assessed his suggestions with regard to how realistically and practically feasible it would be to implement them.

p.3 Gilbert offers a "series of four data presentation/comparison methods":

- (1) ordered listings of the data
- (2) histograms
- (3) box plots
- (4) probability plots

Both (1) and (2) are simple procedures for any spreadsheet program, and should be easy for subcontractors. Items (3) and (4) are "canned" procedures in any statistical software (e.g. StatGraphics, etc.); all subcontractors should use one of the many statistical programs that are commercially available.

p.3 Gilbert recommends a series of six statistical tests:

- (1) hot measurement ("hot spot") test
- (2) Slippage test
- (3) Quantile test
- (4) Wilcoxon Rank Sum (WRS) test
- (5) t-test (if data are normally distributed)
- (6) Gehan test

Of these six, the Gehan test would be the most difficult to perform, and probably would be beyond the capabilities of most subcontractors. This test would require the use of SAS with a specialized SAS code written for the procedure. Additionally, this test has not been studied by Gilbert (or anyone else) with respect to its limitations, power, or robustness. There is some indication that this test may not be so good for comparing samples of different size (which would be a common occurrence in Operable Unit [OU] vs. Background comparisons).

DOCUMENT CLASSIFICATION
REVIEW WAIVER PER
CLASSIFICATION OFFICE

Attachment C

D. M. Smith
September 2, 1993
MAS-006-93
Page 2

Tests (1), (2), and (3) could be easily performed using any spreadsheet. Gilbert suggested that perhaps the UTLs, as calculated in the *1993 Background Geochemical Characterization Report*, could be used for (1), the "hot spot" test. In response to Gilbert's report, we have produced tables giving the 95/95, 95/99, and 99/99 UTLs for all background data. Ralph Lindberg of SMS suggested that we include only the 99/99 results in the 1993 Background Report; all three levels of UTLs would still be available in-house. The 99/99 UTLs would, of course, flag fewer data than the 95/95 UTLs, which would be to our advantage. Tests (4) and (5) would be relatively easy to do using any commercially available statistics program.

Perhaps the most important issues, however, are those only mentioned in passing by Gilbert. Prior to the application of any statistical tests, the data must be reviewed with respect to outliers, units of measurement such as mg/L or ug/L (mixed units could result in strange results), and histograms to examine data distributions. Basic data "cleanup" must precede any data analysis.

The issues of detection limits, result qualifiers, and replacement of non-detects are also topics that need to be addressed before doing any comparisons between OU and Background data. There is good evidence to suggest that, for as much as 50-80% non-detects, simple substitution is really not all that bad (Sanford et al., 1993). Certainly from a pragmatic perspective, simple substitution is the way to go for subcontractors using spreadsheets for much of the data analysis. For higher percentages of non-detects, the methods of Cohen (1961) or Helsel (1990) may be better, although for data sets with multiple detection limits, there is no simple solution.

Particularly in the case of metals (both total and dissolved), the Rocky Flats Environmental Database System data contain multiple detection limits (an average of nine per analyte). The presence of multiple detection limits needs to be considered before selecting a method of replacement for non-detects, and before selecting a method for OU vs. Background comparisons. Plotting of histograms of both the OU and Background data may help to resolve these issues.

jlm

cc:

M. E. Levin *sk Fey MEL*

replacement of non-detects, testing for distribution shape and variance, and conducting appropriate t tests or the WRS test.

As the performance of the Gehan test has not, in my opinion, been adequately determined, I recommend that statistical evaluations and comparisons of its performance with competing tests should be conducted by EG&G at the earliest time. The performance assessments should specifically include data sets that contain one or more nondetects larger than detects. The performance of the Gehan test (or any other test) for this situation has not, to my knowledge, been studied. More generally, future work should include considering how to statistically analyze data sets that contain nondetects that are larger than all detects. ←

Example: We use the RFP data in Figure 1 and a Type I error rate of 0.05. In Figure 7 the ordered background and OU data as well as their Gehan ranks and scores are displayed. Using these scores $[a(R_i)]$ and $m = 10$, $n = 20$, $N = 30$ in the equation for Z , we find that $Z = -0.7376$. Since Z is smaller than 1.645, we conclude that Gehan's test does not indicate the analyte is a PCOC.

Test 6. t test

Purpose: The t test is one of the most widely known statistical tests for testing that the means of two populations are different. When the background and OU data are normally and independently distributed, each distribution has the same variance, and neither data set contains any nondetects, the t test is the preferred test.

Method: The reader is referred to a statistics book for how to conduct a t test, e.g., Snedecor and Cochran (1980, pp. 89-99).

Example: We use the RFP data in Figure 1. However, the t test is not recommended because some OU data are nondetects. The Gehan test should be used instead because nondetects with multiple detection limits are present. If no nondetects were present then the WRS test is appropriate.

Summary Comments for PHASE IV

The tests discussed above have been applied to the data in Figure 1. We found that the HM comparisons identified 2 OU measurements that exceeded the 95% UTL on the 95th percentile. However, the Slippage, Quantile and Gehan tests did not indicate the analyte is a PCOC. The next step is to apply professional judgment, geochemical analyses, and knowledge of RFP (Phase V) to evaluate the validity of the individual measurements and the results of the statistical tests. (These checks supplement the data validity checks made during Phase 2 (data collection/validation.) If uncertainty remains after this evaluation,

Deverly Ramsey
July 30, 1993
Page 6

- The comments and questions offered by statisticians at the workshop indicated they have the skills and knowledge to make significant contributions to the solution of environmental data collection and analysis problems at RFP. However, I perceived that the EG&G statistics group in particular was not well known among the workshop participants. This situation must be corrected. Statisticians should be full team members working closely with others to develop and apply appropriate statistical methods.

**TASK 2: EVALUATE THE APPROPRIATENESS AND APPLICABILITY OF
(A) THE 95% UPPER TOLERANCE LIMIT, AND
(B) ANOVA METHODS AS PROPOSED BY EPA REGION VIII**

We begin by defining what is meant by UTL and ANOVA methods.

95% Upper Tolerance Limit (UTL) on the 95th Percentile

First we define the 95% UTL as it is currently being used at RFP.

Definition: The computed 95% UTL is such that we are 95% confident the UTL is equal to or greater than the true 95th percentile of the population of background measurements.

In other words, the 95% UTL being used at RFP is an upper 95% confidence limit on the 95th percentile of the background distribution. It is possible to compute tolerance limits for other percentiles with other specified degrees of confidence. However, we are concerned here with only the UTL as defined above. The method of computing the 95% UTL on the 95th percentile is given in Appendix C.

ANOVA Procedures

The term "ANOVA" refers to a class of statistical tests and procedures for comparing means or medians of two or more populations. ANOVA procedures include those that are appropriate for normal distributions, such as the t test and one-way analysis of variance, as well as nonparametric tests such as the Wilcoxon Rank Sum (WRS) test and the Kruskal-Wallis test that do not require normally-distributed data. (The latter two tests are nonparametric analogues of the t test and one-way analysis of variance, respectively.) ANOVA methods are very well known by statisticians and practitioners, and are widely used in many fields of application. These methods are discussed in many statistics books, including Sachs (1984) and Snedecor and Cochran (1980).

Agency Positions