PSDDA CLARIFICATION PAPER
7/25/96

SMS TECHNICAL INFORMATION MEMORANDUM

STATISTICAL EVALUATION OF BIOASSAY RESULTS

Prepared by Dr. Teresa Michelsen (Washington Department of Ecology) and Travis C. Shaw (U.S. Army Corps of Engineers) for the PSDDA/SMS agencies.

INTRODUCTION

Sediment bioassays are an integral part of sediment management programs in Washington State. Under the Puget Sound Dredged Disposal Analysis (PSDDA) program and the Washington Sediment Management Standards (SMS), sediment bioassays may be used to determine whether sediments are suitable for open-water disposal, whether sediments require cleanup, and in determining the need for source control to protect sediment quality near a discharge. The interpretation of bioassay results under these programs requires two evaluations:

· A comparison of the response (e.g., mortality) observed in a sample to a threshold value (absolute or relative to a reference response) established by the agencies, and

· An evaluation of whether the adverse effect observed in the sample is statistically significant and greater than the effect observed at a reference station.

The discussion below provides guidance on the determination of statistical significance under these two regulatory programs.

PROBLEM IDENTIFICATION

Regarding the determination of significance, the SMS rule provides that a t-test, $p<0.05$, be used to determine whether the mean of the site station is statistically different from the mean of the reference station. This statement alone does not provide enough detail to ensure that the regulated community and agency staff consistently produce the same results when analyzing data sets. However, WAC 173-204-130(4) provides authority for Ecology to propose technical methods that replace and/or enhance methods provided for in the rule, providing public review is conducted and the decision to use an alternate technical method is documented in the public record.

The PSDDA Phase II Management Plan Report does provide additional guidance on data transformations and statistical tests to be used, discussed in the sections below. The PSDDA guidance further provides a null hypothesis (similar to text in the SMS rule) that the mean of the site and reference stations are not statistically different; however, this null hypothesis is not appropriate for the one-tailed t-tests recommended for use in SMS and PSDDA regulatory programs.

Agency staff and the regulated community have requested further clarification under SMS and PSDDA on the specific form of the t-test to be used, appropriate hypothesis testing, recommended data transformations, and recommended tests for normality and homogeneity of variances. In addition, the agencies have been asked whether Dunnett's test or other alternative tests could be used in place of the t-test. Finally, questions have arisen over the use of multiple reference stations for comparisons under these programs. The discussion below provides guidance on each of these topics.

DISCUSSION AND TECHNICAL BACKGROUND

Hypothesis Testing

In conducting statistical comparisons for sediment management programs, the only concern is whether adverse effects in the sample being tested are greater than adverse effects in a reference sediment (one-tailed hypothesis). The correct null (Ho) and alternate (Hi) hypotheses for comparing the mean response of the test sediment with the reference sediment are:

Ho: Mean test response (e.g., mortality) is less than or equal to the mean reference response at alpha = 0.05

or, Ho: (sample) $\leq$ (reference)

Hi: Mean test response is greater than the mean reference response at alpha = 0.05

or, Hi: (sample) > (reference)

Note that, for the larval bioassays, the alpha level should be increased to 0.10 to account for historically high variances in these tests (see PSDDA clarification paper *Interim Revised Performance Standards for the Sediment Larval Bioassay,* finalized November 10, 1994 and the Draft SMS Technical Information Memorandum *Quality Assurance Guidelines for the Sediment Larval Bioassay* presented with this paper at the 1996 SMARM).

The statement of hypothesis should be revised for effects endpoints where the adverse effect being measured results in a test response lower than the reference (e.g., growth endpoint for *Neanthes).* The correct hypotheses in these cases are:

Ho: Mean test response (e.g., growth rate) is greater than or equal to the reference response at alpha = 0.05.

or, Ho: (sample) $\geq$ (reference)

Hi: Mean test response is less than the reference response at alpha = 0.05.

or, Hi: (sample) < (reference)

In either case, if the null hypothesis is rejected, then we accept the alternate hypothesis that a statistically significant adverse effect is indicated, with a 5% probability of a Type I error (misidentification of an unimpacted station as impacted). If we fail to reject the null hypothesis, we determine that no significant adverse effect has been identified. The direction of the inequality in the statement of hypothesis affects the comparison of the calculated t statistic with the t table value for a given significance level. The proper relationship between the hypothesis and critical region used in decisions about the hypothesis is presented below:

Type of Test	Null Hypothesis	Alt. Hypothesis	Critical Region
1-tailed	u(test) ≤ u(ref)	u(test) > u(ref)	t(calc) > t(table)
1-tailed	u(test) ≥ u(ref)	u(test) < u(ref)	t(calc) < -t(table)

Data Transformations

As noted above, use of the t-test requires that the data are normal and variances are homogeneous. Data derived from bioassay tests are often expressed in terms of percent (mortality or other endpoint). An arcsine-square root transformation may be performed if needed to stabilize the variances and improve the normality of data sets expressed in percent. The arcsine-square root transform is provided below:

y = arcsine (square root of x)

where x is the percentage expressed as a decimal (e.g., 0.80 instead of 80%). This transformation should not be used with bioassay data that are not expressed in percentages, such as growth rate or biomass.

If heterogeneous variances are encountered with biomass data, a log10 transformation may be applied to stabilize the variances. This transformation is typically used for environmental data for which the variance increases as the mean increases (as may be the case for data related to the growth of organisms), and is often successful in making the variance independent of the mean (Sokal and Rohlf, 1969).

Tests for Normality and Homogeneity of Variances

The theoretical basis for the t-test assumes that both samples being tested come from a normal population with equal variances. Violation of these assumptions reduces confidence in the Type I error rate. As a result, tests for homogeneity of variances and normality should be conducted after applying appropriate data transformations and before conducting the t-test.

To test the null hypothesis "the data have been drawn from a normally distributed population", the Wilk-Shapiro statistic (or W test) should be used. A Cochran's test (or F test of variances) should be used to determine whether the variances are homogeneous or heterogeneous.

Form of t-Test

This section affirms that the one-tailed Student's t-test referred to in PSDDA guidance should normally be used in evaluating sediment bioassay data under SMS cleanup and source control programs. This consistency between programs will support cross-comparisons between data sets and reduce the potential for confusion among parties regulated by both programs. Use of the Student's t-test is contingent upon an assumption that the data set is normal and variances are homogeneous. If these conditions are not met following appropriate transformations, the Mann-Whitney test for statistical significance should be used in place of a t-test or approximate t-test.

Use of Multiple Comparison Tests

Several consultants and regulated parties have suggested that the agencies address whether multiple comparison tests (such as ANOVA and Dunnett's) could be used in place of the t-test. Concerns have been raised that the use of multiple pair-wise tests in a single project could increase the Type I error rate for that project. Acceptable Type I error rates have typically been set by the agencies at 5%. However, if a multisample test design is used with a t-test, the Type I error rate increases with each additional station added to the comparison. For example, the null hypothesis tested in a multisample test might be "the mean of sample 1 is the same as the mean of sample 2, which are both the same as the mean of the reference station", or in mathematical terms:

$$\text{Ho: } u(1) = u(2) = u(\text{reference})$$

While a true multisample comparison approach would control Type I error for the above hypothesis at 5%, multiple t-tests used for the same number of comparisons would have a higher Type I error rate (Zar, 1984).

However, this null hypothesis addresses the relationship of each station not only to the reference station, but to other test stations being evaluated. Under PSDDA and SMS bioassay evaluation procedures, this type of evaluation is not conducted. Under PSDDA, dredged material management units (DMMUs) are being evaluated individually for disposal; each may be dredged and disposed of independently of the others. This relationship can be expressed mathematically as:

$$\text{Ho: } u(1) \le u(\text{reference})$$

and

$$\text{Ho: } u(2) \le u(\text{reference})$$

For each comparison, the Type I error rate remains at 5%. Likewise, under SMS, each station is considered individually, and may or may not be compared to the same reference station as another. A hit/no-hit designation is made for each separate site station, and the Type I error for each individual station remains at 5%. Stations are not compared to each other to determine if they are the same or different. Once all stations have been independently tested against an appropriate reference station, the agencies evaluate the number of stations with exceedances and the magnitude of these exceedances to determine the need for cleanup or source control.

An argument was made that the "overall" Type I error rate increases with each additional comparison. However, the agencies are not concerned with this overall error rate, since it is not likely to affect the final regulatory decision. When there is a 5% chance of error at each station, it is clear that as you make hit/no-hit decisions for a number of stations the overall chance of making one incorrect decision somewhere within that area increases. However, one incorrect assignment, or even several, are not likely to make a significant regulatory difference at cleanup sites where 30 or more stations may have been sampled. It is the overall number, pattern, and magnitude of hits that drive cleanup decisions, along with a wide variety of additional types of evidence.

In addition, the agencies are less concerned with a type I error (the chance that a clean station would be designated as dirty) than the type II error (the chance that a dirty station would be designated clean), since it is the latter that determines the power or ability of an agency to detect a contaminated site. At the alpha level set by the SMS, use of the ANOVA/Dunnett's procedure decreases the power, resulting in fewer detection's of contaminated sites. In order to get the same power as the t-test, the alpha level or type I error rate would also have to be increased. Thus it is not obvious that the ANOVA/ Dunnett's procedure offers better performance than the existing method. Therefore, under both programs, the pairwise comparison of the t-test is appropriate to the evaluation procedures that have been adopted.

Use of Multiple Reference Stations

For some projects, samples from multiple reference stations are being collected. This is often done to increase the chances that at least one reference station will meet performance standards, or to collect reference samples representative of different grain size regimes present at the site. In addition, field replicates are sometimes collected to assess sediment heterogeneity or variation due to sampling procedures. Because the SMS rule and PSDDA evaluation procedures were written assuming a single reference station, there has been some uncertainty in how to perform comparisons to reference when there are data for more than one acceptable reference station. The following guidance is provided on assessing bioassay results with multiple reference stations:

· As discussed above, pair-wise comparisons are currently being used in both the PSDDA and SMS programs. Multiple comparison tests that compare the distribution of data at the project location to the distribution of data at a reference area are not appropriate for PSDDA because of the need to treat each individual DMMU separately. The SMS decision process is not structured to allow this type of comparison, and additional development work would need to be done on evaluation procedures if this were contemplated. Therefore, for each site station, a single reference station must be selected for the regulatory comparison.

· If field replicates have been collected at any of the reference stations, the following procedure should be used for statistical tests. Determine the mean response of the lab replicates for each field replicate. Then find the average of the means of all the field replicates and compare this average to the performance standards. This determines whether the station as a whole passes performance standards. An individual field replicate may be excluded or rejected if, in the agency's discretion, it was adversely affected by a sampling, handling, or laboratory problem not representative of environmental conditions at the station.

· For subsequent statistical testing, all field replicate data at a reference station may be pooled, or a representative field replicate may be selected for each station. Pooled data should be analyzed using the nonparametric Mann-Whitney test, since the number of replicate data at the site station will be different from the reference station. In selecting a representative replicate, consideration should be given to the degree of variability as well as absolute response. Treatment of field replicate data for statistical analysis should be discussed in the SAP and approved in advance by the lead agency.

· In cases where grain size varies widely at the site, multiple reference stations may be collected to allow comparison of site stations to the acceptable reference station that most closely matches it in grain size. This is particularly appropriate when the bioassay organism used is known to be affected by grain size (e.g., *Rhepoxynius or Ampelisca).* Reference stations that do not meet performance standards should be eliminated from the evaluation and each site station compared to the remaining reference station with the closest percent fines.

· If grain size is not an issue, the performance of multiple reference stations should be evaluated with respect to all bioassays being conducted, and any reference stations eliminated that do not meet performance standards for all bioassays. If more than one reference station remains that meets all performance standards, a station should be recommended by the project proponent and approved by the lead regulatory agency prior to conducting the statistical analysis. Criteria for selecting an appropriate station could include selection of a station that best represents the overall habitat at the site (e.g., water depth, grain size, TOC), or a station could be selected that is representative of the range of responses (considering both magnitude and variability of the response) seen at the reference area.

· Tables reporting bioassay results should clearly identify which reference station was used for each comparison.

PROPOSED CLARIFICATIONS AND MODIFICATIONS

In summary, the following guidance is provided for the PSDDA and SMS programs:

· A null hypothesis shall be selected that reflects the one-tailed t-test approach and the type of endpoint being evaluated. Appropriate null hypotheses are provided above.

· Bioassay data expressed in percent should be transformed prior to statistical testing using the arcsine-square root transform. This data transformation should not be used for endpoints not expressed in percent (e.g., growth, biomass). A log10 transformation may be used with growth or biomass data.

· Bioassay data should then be tested for normality and homogeneity of variances, using the Wilks-Shapiro test (W test) and Cochran's test (F test for variances), respectively.

· Bioassay data passing both tests should be tested for statistical difference using a one-tailed Student's t-test. Bioassay data failing one or both of these tests should be tested for statistical difference using the nonparametric Mann-Whitney test.

· Multiple comparison tests (e.g., ANOVA, Dunnett's) are not to be used under either SMS or PSDDA.

· If field replicates are collected at reference stations, and/or multiple reference stations are available that pass performance standards, guidance provided above should be followed in using these data for statistical tests.

REFERENCES

Sokal, R.R. and F.J. Rohlf 1969. Biometry: The principles and practice of statistics in biological research. W.H. Freeman and Co, San Francisco, CA.

Zar, J.H. 1984. Biostatistical Analysis, 2nd Ed. Simon & Schuster Co., Englewood Cliffs, NJ.