TECHNICAL GUIDANCE MANUAL FOR HYDROGEOLOGIC INVESTIGATIONS AND GROUND WATER MONITORING

CHAPTER 13
STATISTICS FOR GROUND WATER QUALITY COMPARISON

February 1995

TABLE OF CONTENTS

BASIC STATISTICAL ASSUMPTIONS 13-1
INDEPENDENT SAMPLES 13-2
DETERMINATION OF BACKGROUND DATA SET 13-3
SAMPLING FREQUENCY 13-3
CORRECTIONS FOR SEASONALITY 13-3
STATISTICAL ASSUMPTIONS THAT VARY WITH METHODS 13-3
MINIMUM SAMPLE SIZE 13-4
DETERMINATION OF DISTRIBUTION 13-5
TRANSFORMATION OF DATA TO ACHIEVE NORMAL DISTRIBUTION 13-6
NORMALITY TESTS 13-6
HOMOGENEITY OF VARIANCE 13-7
TREATMENT OF VALUES BELOW THE DETECTION LIMIT (NON-DETECTS . . 13-7
ERRORS 13-8
Comparisonwise Error 13-8
Experimentwise Error 13-8
METHODS 13-9
ANOVA 13-10
TWO-SAMPLE TESTS 13-10
INTERVALS 13-10
Tolerance Intervals 13-11
Prediction Intervals 13-11
Confidence Intervals 13-12
CONTROL CHART METHOD 13-12
STATISTICAL DATA SUBMITTALS 13-12
REFERENCES 13-14

CHAPTER 13

STATISTICS FOR GROUND WATER QUALITY COMPARISONS

The statistical evaluation of data resulting from the chemical analysis of ground water is required by Ohio's regulations for Part B permitted hazardous waste land disposal facilities [OAC 3745-54-97(G)], Part A permitted hazardous waste land disposal facilities [OAC 3745-65-93], and permitted municipal solid waste [OAC 3745-27-10(C)(6)], residual waste [OAC 3745-30-08(C)(5)], and industrial waste landfills [OAC 3745-29-10(C)(6)]. Additionally, statistics can be employed for wastewater facilities (including non-toxic flyash, bottom ash, foundry sand and coal storage piles) and in corrective actions at regulated and unregulated hazardous waste sites. This chapter is intended to address only those statistical evaluations performed for Part B permitted hazardous waste land disposal facilities, permitted municipal solid, residual, and industrial waste landfills, and wastewater facilities. This chapter pertains to Part A permitted hazardous waste land disposal facilities only if the owner/operator has entered into an agreement with the Director of Ohio EPA to utilize methods identified in OAC 3745-54-97 in place of those specified in OAC 3745-65-93. Regarding the methods mandated by OAC 3745-65-93 (i.e., the Student's T-test and the Average Replicate Test), the reader is referred to the U.S. EPA's Technical Enforcement Guidance Document (U.S. EPA, 1986).

Statistical analyses are used to compare the chemical ground water quality of a monitored zone downgradient of the waste management unit with either: 1) a standard set in a permit, 2) the chemical ground water quality from a background (or upgradient) well screened in the same monitored zone and unaffected by facility operations, or 3) historic concentrations from the same well. These comparisons provide reliable determinations as to whether a waste management unit has influenced the quality of the ground water. Statistics can also be used to define the extent of ground water contamination. If no statistically significant difference is observed between a downgradient well and background, the well is statistically considered beyond the plume of contamination.

U.S. EPA has issued the Guidance Document on the Statistical Analysis of Ground Water Monitoring Data at RCRA Facilities, (U.S. EPA, 1989) and an addendum to the same document (U.S. EPA, 1992). The Ohio EPA recommends the methods outlined in the U.S. EPA guidance. Additionally, U.S. EPA has developed a data management tool designed to facilitate the storage, analysis and reporting of data collected through ground water monitoring programs. The statistical portion of the program incorporates the most frequently used statistical methods for ground water monitoring. This tool is called the Ground Water Information Tracking System with Statistical Analysis Capability (GRITS/STAT or GRITS).

This chapter supplements and clarifies the U.S. EPA's guidance with regard to the number of different statistical methods required per facility, independent samples, determination of a background data set, initial year sampling frequency, and corrections for seasonal fluctuations. In addition, fixed and variable requirements or assumptions of the most frequently used statistical methods are discussed. These assumptions include minimum sample size, distribution, variance, treatment of non-detects, and comparison and experimentwise errors. Finally, recommendations for submittals of statistical information are outlined. It is recommended that the U.S. EPA guidance documents be reviewed before reading this chapter.

BASIC STATISTICAL ASSUMPTIONS

Whether starting from scratch with a new database or building on an established database, the first steps in determining whether a statistically significant difference has occurred should be the same. Before choosing the statistical method, an evaluation must be made as to whether the available site-specific data meet a few basic assumptions necessary for a particular test to perform with the greatest accuracy and power, power being the probability that the test will correctly identify contamination when it is present. Assumptions basic to all methods include independence of samples, determination of the background data set, sampling frequency, and corrections for seasonality.

INDEPENDENT SAMPLES

Statistical analysis is based on the assumption that all data points are generated independently of each other. For ground water samples to be independent of each other, enough time must pass between sampling events to ensure that the previously sampled ground water has left the vicinity of the monitoring well and that "new" ground water is being sampled. At the same time, the sampling interval must be short enough to provide "immediate" detection of contamination.

Ohio Hazardous Waste Regulations, OAC Rule 3745-54-97(G)(1), 98(D) and 99(F) and (G), require the collection of independent samples for permitted facilities. Wastewater policy and solid waste regulations do not state specifically that independent samples must be collected, but solid waste regulations do state that the sample number must be appropriate for the statistical method chosen (OAC Rule 3745-27-10(C)(9)). To perform a valid statistical test, including the collection of independent samples, the time interval between sampling events should be based on the average linear velocity of ground water for the zone being sampled. The sampling interval is determined after evaluating the monitored zone's effective porosity, horizontal hydraulic conductivity, hydraulic gradient, and the fate and transport characteristics of potential contaminants. U.S. EPA (1989) outlined the method used to determine the sampling interval necessary to obtain an independent sample.

For wells recovering reasonably quickly, it may be possible to collect multiple independent samples in one scheduled event by purging a well thoroughly, collecting the first independent sample, then repurging the well and collecting the second independent sample, etc., until all necessary samples have been collected. If a regulated entity proposes this option, it must supply data indicating the recharge rate of the well. However, to develop a representative background ground water quality database, temporal and seasonal water level (and parameter) variations must be accounted for. In other words, just because there is adequate recharge does not mean that the complete background data set may be collected at one sampling event.

Generally, replicate samples are unacceptable for statistical analysis because the information they provide indicate only the accuracy of the laboratory, not the ground water quality. When more than one sample is collected for the same parameter from the same bailer or same aliquot of ground water, the samples are considered replicate samples. Replicates are not recommended as they offer less variability than independent samples. If, however, replicates have been obtained as part

of the original background data for hazardous waste Part A Interim Status ground water monitoring, and the Ohio EPA approves the use of this data as background for Part B permitted status statistical analyses, the replicate values should first be averaged and the means used in the statistical analyses.

DETERMINATION OF BACKGROUND DATA SET

For all statistical methods employing the use of pooled background data, the background data set should be such that it reflects naturally occurring changes in hydrogeology. A moving background data set is recommended using data from only the eight most recent sampling events. This will help minimize temporal variability (Sara, 1991). During the initial year of monitoring, only the background or upgradient well(s) should be used as background when comparing up to downgradient values. In subsequent years, the background levels, whether using up to downgradient or intra-well comparisons, should be modified during each monitoring event so that the eight most recent values are used. When intra-well comparisons are being performed (e.g., control chart) the background database must be from an uncontaminated well.

SAMPLING FREQUENCY

Minimum sampling frequency shall be consistent with the appropriate statistical method chosen. The Ohio hazardous waste rules require ***at least*** semi-annual sampling for detection monitoring (OAC Rule 3745-54-98(D)), while solid waste rules (OAC Rule 3745-27-10(D)(5)) require ***at least*** semi-annual monitoring for the initial year. However, to determine initial background values, sampling must be appropriate for the statistical method chosen (OAC Rule 3745-54-97(G)). The hazardous or solid waste permit should outline the sampling frequency necessary to achieve the minimum sample size, which may oblige the permittee to sample ***more than*** what appears to be regulatorily required.

CORRECTIONS FOR SEASONALITY

***Only when strong empirical evidence is present to indicate that seasonality exists should corrections for seasonality be made*.** If seasonality is present in a data set, Ohio hazardous and solid waste regulations require procedures be used to control or correct for it. The following methods are available: 1) Two-way ANOVA, which attributes variations to seasonal as well as spatial differences, 2) U.S. EPA (1989) provided a simple method for calculating the monthly or quarterly effects attributable to seasonal variations, and 3) Seasonal Kendal Test (Gibbons, 1994). Tolerance intervals tend to be self-correcting for seasonality since seasonal effects show up in both upgradient and downgradient wells.

STATISTICAL ASSUMPTIONS THAT VARY WITH METHODS

Once the basic assumptions have been met or the data set has been transformed to meet the basic assumptions, a statistical method may be chosen. The next step in making the choice is to determine the best fit between the site-specific data available and the specific assumptions that allow each method to perform with the greatest accuracy and power. Assumptions that vary

between methods include minimum sample size, determination of distribution, homogeneity of variance, treatment of non-detect parameter levels, and experimentwise and comparisonwise errors. Another factor to take into account for each of the following assumptions is, as the-site specific database changes over time, it may be necessary to change to a different statistical method.

MINIMUM SAMPLE SIZE

Statistical methods must employ a certain minimum sample size to yield both statistically and hydrogeologically valid results. The term "sample size" reflects the number of observations for each parameter taken per well. Minimum sample size requirements are provided below for common methods. Although it is possible to perform the calculations and obtain mathematically valid results using fewer independent samples for some statistical methods, the minimum sizes are designed to yield hydrogeologically valid results. Permit applicants should be aware that during the initial year of monitoring, the ***minimum*** sampling frequency is semi-annual and that more frequent sampling may be necessary, depending on the statistical method chosen, the hydrogeologic environment, and the objectives of the study. Over time, the minimum sample size may be met, allowing the use of a different statistical method.

- ***ANOVA (Parametric or normally distributed)***: The minimum sample size for analysis of variance (ANOVA) should result in at least ten (10) degrees of freedom for error (or within groups) variance. The number of degrees of freedom is the number of ways the sample may be changed without having to produce any change in the constraining factors. For ANOVA, the total number of wells should be equal to or greater than two (2). The degrees of freedom are calculated using the equation N-p, where:

  N = total number of observations; and
  p = total number of wells.

  Example: Quarterly sampling from four (4) wells will give twelve (12) degrees of freedom (N-p=16-4=12).

  U.S. EPA (1989) recommends that N - p be greater than 5 with $p \geq 2$, and with at least three (3) observations per well. However, if p (the number of wells) = 2, then the test is essentially equal to the t-test. Care should be exercised with small sample sizes because they may lead to false positives (statistical test indicating contamination when none exists) and false negatives (statistical test indicating no contamination exists when, in fact, there is contamination). It is better to recommend a minimum of 10 degrees of freedom, with the number of wells being equal to/greater than 2 ($N - p \geq 10$, $p \geq 2$).

- ***ANOVA (Non-parametric or non-normal distribution):*** The minimum sample size required for Non-parametric ANOVA (ANOVA based on ranks) is at least three (3) wells with at least three (3) observations in each well (N-p=9-3=6). Ohio's solid waste regulations require that a minimum of four (4) independent samples be taken from each well during the initial 180 days of monitoring, so there will be at least four (4) independent observations per well initially.

- ***Tolerance Interval**:* The minimum sample size for constructing a Tolerance Interval is at least eight (8) independent background observations from each well. Tolerance interval calculations can be made with as few as three (3) observations; however, this would result in a large upper tolerance limit due to the increased tolerance factor (K) associated with smaller sample sizes.

- ***Prediction Intervals***: The minimum sample size for constructing prediction intervals should result in at least ten (10) degrees of freedom. Degrees of freedom for tolerance intervals are calculated using the equation n - 1, where n = number of observations in the background data (so at least 11 independent samples are needed). While eight (8) samples may be used, ten (10) or more are recommended. The prediction limit should be recalculated at least annually to update the background for upgradient changes.

- ***Control Charts***: The minimum sample size required for control charts is eight (8) observations in ***each*** well. This requires that eight ***independent*** samples be collected over a one year period for each well.

DETERMINATION OF DISTRIBUTION

Normal distribution is based on the Central Limit Theorem, which states that sums and averages of random variables tend to be normally distributed. Normality deals with average behavior and average variability of behavior. Distribution is important in ground water monitoring because determination of population characteristics are being made based on limited information contained in a set of data. The most common example of normal distribution is the bell-shaped curve. The assumption that all samples are independent is required.

If data is normally distributed, parametric methods of analysis may be applied. When data is neither normally or log-normally distributed, non-parametric methods are most commonly be used. Parametric methods may be applied to non-normally distributed data and non-parametric tests may be applied to normal data. Non-parametric methods are more efficient, or powerful, when data is normally distributed than parametric methods when if data is not normal.

Skewness measures the degree of symmetry of the sample distribution. Normal data has a skewness coefficient of 0. If the data is neither normally or log-normally distributed, is positively skewed, and an indication exists that the distribution is skewed to the right, the statistician should proceed as if the data was log-normal using non-parametric tests or try another type of transformation. Parametric tests lose power when the skewness coefficient is > 1.0. If the same case exists and the data is negatively skewed, indicating a shift to the left, Cohen's adjustment for non-detect data or non-parametric tests should be used. If the skewness is equal to zero, a perfect bell-shaped curve exists.

Kurtosis is the area under the distribution curve. The greater the spread of the data distribution, the lower and broader the peak of the distribution curve will be. Kurtosis measures the degree of peakedness of the sample distribution. It is measured relative to the normal distribution curve. Normal kurtosis is = 0. As the kurtosis coefficient becomes greater, the peakedness decreases and the curve spreads out, indicating a broader distribution of values. As the kurtosis coefficient

decreases (below 0), the peak of the distribution curve becomes higher and more peaked, indicating that the distribution is centered around a narrow range of values (Ott, 1977).

TRANSFORMATION OF DATA TO ACHIEVE NORMAL DISTRIBUTION

It is critical in any statistical evaluation to determine whether the data is normally distributed. If the original data does not have a normal distribution, transformations should then be attempted to achieve normality, as most statistical tests have the underlying assumption that the data has a normal distribution. The most accepted method is log-transformation. Log-transformations are most useful when the ratio of the largest to the smallest value in a data set is greater than 20. While U.S. EPA (1992) gives log transformations as the default method, the Ohio EPA recommends both original and log-transformed data be evaluated to find the best fit. Transformations other than logging the data, such square root, reciprocal and cube root, while less standard, may also be applied. If normality cannot be achieved, Non-parametric tests should be used.

NORMALITY TESTS

Many different tests are available to determine if data is normally distributed. The following discussion will describe the minimum requirements necessary to perform the eight most common normality tests, Shapiro Wilk, Shapiro Francia, Chi Squared, Coefficient of Variation, Probability Plots, Kolomorgorov-Smirnov, Skewness Coefficient, and Kurtosis. All of these tests are available in GRITS.

- ***Shapiro Wilk:*** Shapiro Wilk is considered the best numerical test of normality. It is the most powerful for detecting departures from normality in the tails of a sample distribution. It is useful for sample sizes ranging from 3 to 50 and data must be log-transformed before performing the test

- ***Shapiro Francia:*** While Shapiro Francia has the same benefits as Shapiro Wilk, it is best used for sample sizes larger than 50. It also should be used on log-transformed data.

- ***Chi Squared (X2 = Mean^2)*:** Chi Squared is not the most powerful test, as it does not indicate ***how*** the data is not normal. If the departure from normality is in the tails instead of the middle of the data, Chi Squared may not register it as significant.

- ***Coefficient of Variation*:** This test is easy to calculate: if S/X > 1 (S/X= Standard Deviation/ Mean), the distribution is not normal. While this test is good for small sample sizes, it is not a reliable indicator of model appropriateness. Its true purpose is to estimate skewness, not normality.

- ***Probability Plots/Correlation Coefficient:*** This method measures the strength of the linear relationship between two variables, normal distribution, and the test group. The test group value should, if normal, fall in a straight line with the proportion of observations less than or equal to each observed value. The correlation coefficient is = 1.) If no relationship exists between the two groups, the correlation coefficient = 0. This test is a good indicator of skewness and the presence of outliers.

- ***Kolomorgorov-Smirnov Test (Cheeney, 1983*)**: The Kolomorgorov-Smirnov test graphically evaluates the fit of the distribution (the goodness of fit).

- ***Kurtosis*:** Kurtosis measures the thickness of the tails in a distribution. Normal distribution has a kurtosis of 0.

- ***Skewness Coefficient*:** The skewness coefficient measures the degree of symmetry in the distribution. Normal distribution has a coefficient of 0. Parametric tests lose power when the skewness coefficient is > 1.0.

HOMOGENEITY OF VARIANCE

Variance is the variation or skewness between the parameter levels in the wells being compared, typically downgradient and background wells. It estimates the amount of spread or variability of the data. Homogeneity of variance assumes that all the wells being compared have the same skewness or variance. If this assumption is not true, if the wells do not have close to the same distribution, then the ability of a method to detect differences between the group means is reduced. Methods such as ANOVA assume homogeneity of variance; however, in reality this situation almost never exists. Limitations and uses are described below for the two most commonly used methods of evaluating homogeneity of variance: Box Plots and Levene's Test. Box Plots are covered briefly in Chapter 12. Both tools are available in GRITS.

- ***Box Plots*** provide a quick way to visualize the spread of data using a graphical display. If the longest and shortest box lengths differ by a ratio of greater than 3, Levene's test should be performed for homogeneity of different groups. Box plots are a pre-requisite to performing Parametric ANOVA.

- ***Levene's Test*** evaluates the homogeneity of variance between compliance wells and ***pooled*** background wells. It may be used with either a normal or non-normal data distribution. Variances of different wells are assumed to be approximately equal. Central to the test is the calculation of the F-statistic, which is the detection of differences among group means. If the F-statistic is not significant, variances are approximately equal. If the F statistic is significant, the groups do not have equal variance, and the non-parametric Kruskal Wallis test should be employed. Levene’s test may be used as a One-way Parametric ANOVA test.

TREATMENT OF VALUES BELOW THE DETECTION LIMIT (NON-DETECTS)

If there are less than or equal to 15% non-detects in the total number of measurements analyzed for each parameter, the value of each non-detect should be substituted with one-half the practical quantification limit (PQL) for the parameter (Gibbons, 1992). If 15 to 50% non-detects are present, either use a non-parametric ANOVA test or adjust the original data and proceed with a parametric test. If a method involving intervals is being employed, use either Cohen's or Aitcheson's Adjustments. Cohen's Adjustment assumes that the non-detects indicate a low but positive concentration. Aitcheson's adjustment assumes that the non-detects represent a true zero

concentration. When 50 to 90 % of the results are non-detects, switch to non-parametric rank methods or use Poisson-based non-parametric prediction or tolerance intervals. (However, Wilcoxin Rank Sum is most powerful at 15% <ND < 50%.) If there are 90-99% non-detects, use the Poisson-based non-parametric prediction or tolerance intervals. If 100% of the samples are non-detect, do not employ statistics.

In order to choose between Cohen's and Aitcheson's adjustments, consideration should be given as to whether contamination is present. If so, Cohen's Adjustment may be preferred. Additionally, parameter characteristics should be considered. Non-detects for naturally occurring parameters may indicate a low but positive concentration. In this case, Cohen's Adjustment should be used. If the parameter is a common lab contaminant, Aitcheson's Adjustment should be chosen.

ERRORS

Two types of errors, experimentwise and comparisonwise, are common when performing statistical analysis. Comparisonwise error is based on the false positive rate associated with a single well comparison. Experimentwise error is based upon the total number of statistical tests being performed.

Comparisonwise Error

There are two types of statistical comparisons, paired and multiple. A paired comparison compares one upgradient well with one downgradient well. Multiple comparisons exist when downgradient data is pooled and then compared with upgradient data. Examples of paired comparisons using parametric tests would be the Student t and Average Replicate tests. The Wilcoxin Rank Sum test is a paired comparison test that works well when the data is not normal (non-parametric) and when a large number (even greater than 50%) of non-detects are present, as it is based on ranks rather than actual concentrations.

Multiple comparisons arise where there are multiple compliance wells and multiple parameters per well. If one test is run in every well for each parameter, a large number of tests will be needed. The larger the number of tests, the greater the chance of error. In particular, with increasing comparisons, a corresponding increase in the false positive error rate occurs (i.e., for every 100 tests, 5% of the tests should be expected to be false positives).

Experimentwise Error

Statistics are based on the null hypothesis, that is that there is no real difference between the value of a sample in the population sampled and the hypothesized value of the sample; in other words, the assumption is that there will be no contamination. There are two types of errors associated with accepting this hypothesis:

- ***Type 1 or False Positives ("Hanging the Innocent"):*** Rejection of a true null hypothesis. If there is indeed no contamination (a true null hypothesis) and the statistical test indicates that contamination exists, a false positive has occurred.

- ***Type 2 or False Negatives ("Freeing the Guilty"):*** Acceptance of a false null hypothesis. If the null hypothesis is false, meaning there is indeed contamination, and the statistical test indicates no contamination, a false negative has occurred.

The goal is to minimize both Type I and Type 2 errors. When using ANOVA, the experimentwise error level (Type I, or false positive error) should be less than or equal to five (5) percent for each testing period. This means that, for the upgradient to downgradient well group comparisons, the Type 1 (alpha) error level should be less than or equal to five (5) percent. For individual well comparisons, the error level (Type I error) should be less than or equal to one (1) percent. If there are more than five downgradient wells, each individual well comparison must be made at the one percent error level. The experimentwise five percent error should not be divided by the number of downgradient wells because this may result in an individual error level of less than one percent. If the false positive rate increases, the statistical test will become less powerful.

The power of a statistical test indicates its sensitivity and the probability that the test will indicate actual contamination. The goal is to have the power be as high as possible. If the basic assumptions of each test are met, the maximum power will be employed. The power to detect differences tends to increase as the alternative mean moves farther from the null hypothesis (or as the value from an individual well moves further from the mean of the background database).

METHODS

The following section provides a very general discussion of specific uses and problems associated with the most common statistical methods not addressed in U.S. EPA (1989). This document, along with its Addendum, provide a further discussion of each method as well as sample calculations.

Each constituent that must be statistically analyzed may require a different statistical method; for example, inorganic parameters such as chloride or certain metals may be detected during every sampling event with very few non-detects, thereby allowing analysis using a parametric method. Other parameters with a large percentage of non-detect values may require a non-parametric approach. All statistical assumptions (e.g., data distribution, normality, variance) must be tested and validated upon the designated data set prior performing the statistical test. The statistical test must be appropriate for each constituent's data set.

ANOVA

ANOVA (parametric one way) compares background values to several compliance wells at once. It is a powerful test when only a small amount of data is available. It is also recommended when ground water velocity is higher than average due to the fact that simultaneous testing of multiple compliance wells requires many independent samples. The purpose of ANOVA is to assess whether the average concentration (mean) at any compliance well is significantly higher than mean background level. ***Two-way ANOVA*** is used to correct for seasonal variations. Both are available in GRITS/STAT.

Problems associated with ANOVA are: 1) pooling of downgradient data results in a slower detection of a release; 2) ANOVA commonly indicates a statistically significant difference when small consistent spatial differences in ground water geochemistry exist (which often occurs across a waste disposal facility); 3) both parametric and non-parametric ANOVA assume homogeneity of variance, a condition that almost never exists in reality; 4) ANOVA does not adjust for multiple comparisons due to multiple constituents which results in an increased rate of false positives and negatives (Gibbons, 1993); and 5) ANOVA is not good at detecting a narrow plume that may affect only 1 out of 10 or 20 monitoring wells.

ANOVA (non-parametric) analysis includes the Kruskal Wallis and Wilcoxin Rank Sum tests, which are based on ranking the order of the combined background and compliance data, followed by a comparison of the relative rank of each group. This test is very powerful when the distribution of the data set is non-normal. Non-parametric ANOVA may be used whenever parametric ANOVA is appropriate. Non-parametric methods are more powerful if data is normally distributed than parametric methods are if data is not normal. Many of the same problems exist with non-parametric ANOVA as did with parametric ANOVA (i.e. assumption of homogeneity of variance). In addition, the non-parametric method requires more observations than the parametric ANOVA. It, too, is available in GRITS/STAT.

TWO-SAMPLE TESTS

Wilcoxin Rank Sum and T-Test are examples of two-sample tests. The Wilcoxin Rank Sum test is based on a comparison of 1 background to 1 compliance well, while the T-test compares a pooled background data set with 1 compliance well. The Wilcoxin Rank Sum test does not require normality of distribution, while the T-test does. The power of the two tests is greatest when the percent of non-detects is between 15 and 50%. Both are available in GRITS/STAT.

INTERVALS

Due to the fact that the width of tolerance, prediction and confidence intervals (limits) may be very different for the same data sat, it is important to distinguish the purpose for using intervals before choosing the type of interval to be used. Typically, based on the same data, tolerance intervals (TI) will have the widest limits followed by prediction (PI), then confidence (CI). (TI > PI > CI).

Tolerance and prediction intervals are calculated by comparing background or intra-well data against downgradient sample values. Confidence intervals are computed by comparing downgradient values against a fixed standard such as an ACL or MCL. When using tolerance and prediction intervals, significant evidence of contamination is indicated by any value from either a downgradient or compliance well exceeding the upper 95% tolerance or prediction limit. Statistical evidence of contamination using confidence intervals is indicated when the lower limit of the interval is above the fixed standard. Parametric tolerance and prediction intervals must be normally or log-normally distributed. Typically, log-normal data is used.

The important difference between tolerance and prediction intervals is the definition of "population" or “k”. Tolerance intervals assume a 95% confidence level of including ***a specified portion*** of the entire distribution of measurements from which the background data were drawn. For prediction

intervals, the 95% confidence level is related to containing all of the next k measurements, where k is relatively small. Tolerance intervals are used when comparing data with MCLs, while confidence intervals are used most often with ACLs. This is due to the fact that tolerance intervals are more stringent and should be used when quantifiable health-based risk levels are available.

The following is a discussion of the uses and problems associated with tolerance intervals performed on background and compliance limits, parametric and poisson-based prediction intervals, and parametric and poisson-based confidence intervals. All the intervals are available on GRITS/STAT with the exception of the Poisson-Based Intervals.

Tolerance Intervals

- ***Tolerance Interval On Background*** are used in detection ground water monitoring when comparing the upper limit of uncontaminated background data to individual compliance points. If any single compliance well exceeds the upper 95% tolerance limit, there is significant evidence of contamination.

- ***Tolerance Interval On Compliance Limits (MCL or ACL)*** are Poisson-Based Intervals. This test is most powerful when non-detects are greater than 50%, computed on either parametric or non- parametric data. This method is used in compliance ground water monitoring when comparing the upper limit of a compliance well's data and the fixed Ground Water Protection Limit set in the permit, either the MCL, risk-based limit, or ACL. If the tolerance limit exceeds the standard, significant evidence exists that more than 5% of all compliance well values exceed the tolerance limit. This test is also used in intra-well comparisons when comparing present data with upper limits derived from past data from the same well.

Prediction Intervals

- ***Parametric Prediction Intervals*** are the most powerful when non-detects comprise less than 50% of the data set. The main problem associated with this method is the determination of "K". Cameron (1995) defined K as the number of samples collected from one well during the time between the last sampling event and the next time the data must be statistically analyzed. If statistics are required annually and four sampling events occurred during the year, then K = 4. Parametric prediction intervals are the most powerful when performed on parameters naturally detected in ground water such as inorganics or geochemical parameters. This method is especially useful when only a few compliance data points are available, as in the early stages of detection monitoring. One of the benefits of this method is that it effectively limits the false positive rates without sacrificing false negative rates.

- ***Poisson-Based Prediction Limits (Non-Parametric)*** use original measurements, not ranks. The Poisson intervals are useful for rare event data, when a large percentage of the values are non-detect. They are commonly used for intrawell comparisons when computing the upper interval limit from past data to predict expected values of future well samples.

Confidence Intervals

- ***Parametric Confidence Intervals*** indicate the proportion, over time, that repeated random intervals will cover the desired parameter value. Normal or log-normal data is necessary to perform this test. Parametric confidence intervals are useful if none of the on-site wells are truly upgradient. This method should ***only*** be used in compliance monitoring when comparing the value from a well to a fixed limit (most commonly ACL) derived from **on-site** background data. However, it is not the preferred method for statistical analysis of ground water when the comparison is being made to a health-based standard because this method only estimates the approximate level of the true concentration range. It should ***never*** be compared to a MCL.

- ***Poisson-Based Confidence Intervals (Non-Parametric)***, while requiring more compliance observations than its parametric counterpart, does not always need as large a background data set as the parametric confidence interval. Again, it should ***only*** be used when comparing a value to a fixed limit derived from site-specific data.

CONTROL CHART METHOD

The control chart method, mentioned specifically in the hazardous waste (permitted) and solid waste regulations, is based on normally distributed data but does not require it. As mentioned above, control charts require eight independent samples over a one (1) year period from each well. Unlike the other methods, which compare sample results from upgradient to downgradient wells, the control chart method evaluates the change in concentration of a parameter ***in a single well*** over time. The control chart method should be used only for initially ***uncontaminated*** wells. Before performing this method, a demonstration must be made, using another statistical method, to provide evidence that no contamination is present in the well. Caution is needed when approving this method for sites already in operation where contamination may have occurred previously (U.S. EPA, 1992a).

The control chart method is useful for analyzing inorganics or geochemical parameters having few non-detects. Use of this method avoids problems arising from concentration differences due to spatial or hydrogeologic variability. It is available in GRITS/STAT.

STATISTICAL DATA SUBMITTALS

To facilitate correct interpretation of the statistical data, by both Ohio EPA and the regulated entity, the Agency recommends that certain information be submitted with any statistical evaluations. Evidence should be provided that the statistical methods used:

- were as specified in the permit, if applicable.

- were conducted separately for each constituent.

- were completed within the time frame specified in the permit, if applicable.

- were appropriate for the distribution of the constituents.
- were done at a Type 1 error level of .01 for each testing period for individual well comparisons; and .05 for multiple comparisons (with the exception of tolerance, confidence, prediction intervals and control charts).
- accounted for data below the limit of detection as specified in the permit, if applicable.
- included procedures to control or correct for seasonal, spatial and temporal variability, as necessary.

Example calculations for each statistical method should be provided.

REFERENCES

Cameron, K. 1995. SAIC. Statistical Consultant to U.S. EPA. Personal Communication.

Cheeney, R.F. 1983. Statistical Methods in Geology. George Allen and Unwin, Ltd., London, U.K.

Davis, J.C. 1986. Statistics and Data Analysis in Geology. John Wiley and Sons Publishers, New York.

Gibbons, R.D. 1992. Paper prepared for Waste Management Inc.: Practical Quantitation Limits and Method Detection Limits for Volatile Organic Compounds at WMI Environmental Monitoring Laboratory.

Gibbons, R.D. 1993. A Review of the U.S. EPA Addendum to the Statistical Guidance Document. A paper prepared for Waste Management, Inc.

Gibbons, R.D. 1994. Statistical Methods For Groundwater Monitoring, John Wiley and Sons, New York.

Gilbert, R.O. 1987. Statistical Methods for Environmental Pollution Monitoring. Van Nostrand Reinhold, New York.

Ott, L. 1977. An Introduction to Statistical Methods and Data Analysis. Wadsworth Publishing Company. California.

Sara, M. and R.D. Gibbons. 1991. Organization and Analysis of Water Quality Data. In: D.M. Neilsen (editor), Practical Handbook of Ground Water Monitoring. Lewis Publishers, Inc. Chelsea, Michigan. pp.557-583.

U.S. EPA. 1986. RCRA Ground Water Monitoring Technical Enforcement Guidance Document. Office of Waste Programs Enforcement. Office of Solid Waste and Emergency Response. OSWER-9950.1. Washington, D.C.

U.S. EPA. 1989. Guidance Document on the Statistical Analysis of Ground Water Monitoring Data at RCRA Facilities. Interim Final Guidance. Office of Solid Waste Management Division. Washington, D.C.

U.S. EPA. 1992a. Statistical Analysis of Ground Water Monitoring Data at RCRA Facilities. Addendum to Interim Final Guidance. Office of Solid Waste, Waste Management Division. Washington, D.C.

U.S. EPA. 1992b. User Documentation/ A Ground Water Information Tracking System with Statistical Analysis Capability/ Version 4.2. Document EPA/625/11-9/002. Office of Solid Waste Management, Office of Research and Development. Washington, D.C.

U.S. EPA. 1994. Workbook for the Statistical Training Course for Ground Water Monitoring Data Analysis. EPA Document EPA530-R-93-003. Office of Solid Waste and Emergency Response. Washington, D.C.