Nov 2002

Rec 11/20/02 Reference

LIBRARY COPY

STATISTICAL ANALYSIS OF GROUND-WATER MONITORING DATA AT RCRA FACILITIES

ADDENDUM TO INTERIM FINAL GUIDANCE

OFFICE OF SOLID WASTE
PERMITS AND STATE PROGRAMS DIVISION
U.S. ENVIRONMENTAL PROTECTION AGENCY
401 M STREET, S.W.
WASHINGTON, D.C. 20460

JULY 1992

9570

DISCLAIMER

This document is intended to assist Regional and State personnel in evaluating ground-water monitoring data from RCRA facilities. Conformance with this guidance is expected to result in statistical methods and sampling procedures that meet the regulatory standard of protecting human health and the environment. However, EPA will not in all cases limit its approval of statistical methods and sampling procedures to those that comport with the guidance set forth herein. This guidance is not a regulation (i.e., it does not establish a standard of conduct which has the force of law) and should not be used as such. Regional and State personnel should exercise their discretion in using this guidance document as well as other relevant information in choosing a statistical method and sampling procedure that meet the regulatory requirements for evaluating ground-water monitoring data from RCRA facilities.

This document has been reviewed by the Office of Solid Waste, U.S. Environmental Protection Agency, Washington, D.C., and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the U.S. Environmental Protection Agency, nor does mention of trade names, commercial products, or publications constitute endorsement or recommendation for use.

CONTENTS

1. CHECKING ASSUMPTIONS FOR STATISTICAL PROCEDURES 1

1.1 Normality of Data 1

1.1.1 Interim Final Guidance Methods for Checking Normality 3

1.1.2 Probability Plots 5

1.1.3 Coefficient of Skewness 8

1.1.4 The Shapiro-Wilk Test of Normality (n 50) 9

1.1.5 The Shapiro-Francia Test of Normality (n>50) 12

1.1.6 The Probability Plot Correlation Coefficient 13

1.2 Testing for Homogeneity of Variance 20

1.2.1 Box Plots 20

1.2.2 Levene's Test 23

2. RECOMMENDATIONS FOR HANDLING NONDETECTS 25

2.1 Nondetects in ANOVA Procedures 26

2.2 Nondetects in Statistical Intervals 27

2.2.1 Censored and Detects-Only Probability Plots 28

2.2.2 Aitchison's Adjustment 33

2.2.3 More Than 50% Nondetects 34

2.2.4 Poisson Prediction Limits 35

2.2.5 Poisson Tolerance Limits 38

3. NON-PARAMETRIC COMPARISON OF COMPLIANCE DATA TO BACKGROUND ... 41

3.1 Kruskal-Wallis Test 41

3.1.1 Adjusting for Tied Observations 42

3.2 Wilcoxon Rank-Sum Test for Two Groups 45

3.2.1 Handling Ties in the Wilcoxon Test 48

4. STATISTICAL INTERVALS: CONFIDENCE, TOLERANCE, AND PREDICTION 49

4.1 Tolerance Intervals 51

4.1.1 Non-parametric Tolerance Intervals 54

4.2 Prediction Intervals 56

4.2.1 Non-parametric Prediction Intervals 59

4.3 Confidence Intervals 60

5. STRATEGIES FOR MULTIPLE COMPARISONS 62

5.1 Background of Problem 62

5.2 Possible Strategies 67

5.2.1 Parametric and Non-parametric ANOVA 67

5.2.2 Retesting with Parametric Intervals 67

5.2.3 Retesting with Non-parametric Intervals 71

6. OTHER TOPICS 75

6.1 Control Charts 75

6.2 Outlier Testing 80

ACKNOWLEDGMENT

This document was developed by EPA's Office of Solid Waste under the direction of Mr. James R. Brown of the Permits and State Programs Division. The Addendum was prepared by the joint efforts of Mr. James R. Brown and Kirk M. Cameron, Ph.D., Senior Statistician at Science Applications International Corporation (SAIC). SAIC provided technical support in developing this document under EPA Contract No. 68-W0-0025. Other SAIC staff who assisted in the preparation of the Addendum include Mr. Robert D. Aaron, Statistician.

Draft

STATISTICAL ANALYSIS OF GROUND-WATER MONITORING DATA AT RCRA FACILITIES

ADDENDUM TO INTERIM FINAL GUIDANCE

JULY 1992

This Addendum offers a series of recommendations and updated advice concerning the Interim Final Guidance document for statistical analysis of ground-water monitoring data. Some procedures in the original guidance are replaced by alternative methods that reflect more current thinking within the statistics profession. In other cases, further clarification is offered for currently recommended techniques to answer questions and address public comments that EPA has received both formally and informally since the Interim Final Guidance was published.

1. CHECKING ASSUMPTIONS FOR STATISTICAL PROCEDURES

Because any statistical or mathematical model of actual data is an approximation of reality, all statistical tests and procedures require certain assumptions for the methods to be used correctly and for the results to have a proper interpretation. Two key assumptions addressed in the Interim Guidance concern the distributional properties of the data and the need for equal variances among subgroups of the measurements. In the Addendum, new techniques are outlined for testing both assumptions that offer distinct advantages over the methods in the Interim Final Guidance.

1.1 NORMALITY OF DATA

Most statistical tests assume that the data come from a Normal distribution. Its density function is the familiar bell-shaped curve. The Normal distribution is the assumed underlying model for such procedures as parametric analysis of variance (ANOVA), t-tests, tolerance intervals, and prediction intervals for future observations. Failure of the data to follow a Normal distribution at least approximately is not always a disaster, but can lead to false conclusions if the data really follow a more skewed distribution like the Lognormal. This is because the extreme tail

behavior of a data distribution is often the most critical factor in deciding whether to apply a statistical test based on the assumption of Normality.

The Interim Final Guidance suggests that one begin by assuming that the original data are Normal prior to testing the distributional assumptions. If the statistical test rejects the model of Normality, the data can be tested for Lognormality instead by taking the natural logarithm of each observation and repeating the test. If the original data are Lognormal, taking the natural logarithm of the observations will result in data that are Normal. As a consequence, tests for Normality can also be used to test for Lognormality by applying the tests to the logarithms of the data.

Unfortunately, all of the available tests for Normality do at best a fair job of rejecting non-Normal data when the sample size is small (say less than 20 to 30 observations). That is, the tests do not exhibit high degrees of statistical power. As such, small samples of untransformed Lognormal data can be accepted by a test of Normality even though the skewness of the data may lead to poor statistical conclusions later. EPA's experience with environmental concentration data, and ground-water data in particular, suggests that a Lognormal distribution is generally more appropriate as a default statistical model than the Normal distribution, a conclusion shared by researchers at the United States Geological Survey (USGS, Dennis Helsel, personal communication, 1991). There also appears to be a plausible physical explanation as to why pollutant concentrations so often seem to follow a Lognormal pattern (Ott, 1990). In Ott's model, pollutant sources are randomly diluted in a multiplicative fashion through repeated dilution and mixing with volumes of uncontaminated air or water, depending on the surrounding medium. Such random and repeated dilution of pollutant concentrations can lead mathematically to a Lognormal distribution.

Because the Lognormal distribution appears to be a better default statistical model than the Normal distribution for most ground-water data, it is recommended that all data first be logged prior to checking distributional assumptions. McBean and Rovers (1992) have noted that "[s]upport for the lognormal distribution in many applications also arises from the shape of the distribution, namely constrained on the low side and unconstrained on the high side.... The logarithmic transform acts to suppress the outliers so that the mean is a much better representation of the central tendency of the sample data."

Transformation to the logarithmic scale is not done to make "large numbers look smaller." Performing a logarithmic or other monotonic transformation preserves the basic ordering within a

data set, so that the data are merely rescaled with a different set of units. Just as the physical difference between 80 Fahrenheit and 30 Fahrenheit does not change if the temperatures are rescaled or transformed to the numerically lower Celsius scale, so too the basic statistical relationships between data measurements remain the same whether or not the log transformation is applied. What does change is that the logarithms of Lognormally distributed data are more nearly Normal in character, thus satisfying a key assumption of many statistical procedures. Because of this fact, the same tests used to check Normality, if run on the logged data, become tests for Lognormality.

If the assumption of Lognormality is not rejected, further statistical analyses should be performed on the logged observations, not the original data. If the Lognormal distribution is rejected by a statistical test, one can either test the Normality of the original data, if it was not already done, or use a non-parametric technique on the ranks of the observations.

If no data are initially available to test the distributional assumptions, "referencing" may be employed to justify the use of, say, a Normal or Lognormal assumption in developing a statistical testing regimen at a particular site. "Referencing" involves the use of historical data or data from sites in similar hydrogeologic settings to justify the assumptions applied to currently planned statistical tests. These initial assumptions must be checked when data from the site become available, using the procedures described in this Addendum. Subsequent changes to the initial assumptions should be made if formal testing contradicts the initial hypothesis.

1.1.1 Interim Final Guidance Methods for Checking Normality

The Interim Final Guidance outlines three different methods for checking Normality: the Coefficient-of-Variation (CV) test, Probability Plots, and the Chi-squared test. Of these three, only Probability Plots are recommended within this Addendum. The Coefficient-of-Variation and the Chi-squared test each have potential problems that can be remedied by using alternative tests. These alternatives include the Coefficient of Skewness, the Shapiro-Wilk test, the Shapiro-Francia test, and the Probability Plot Correlation Coefficient.

The Coefficient-of-Variation is recommended within the Interim Guidance because it is easy to calculate and is amenable to small sample sizes. To ensure that a Normal model which predicts a significant fraction of negative concentration values is not fitted to positive data, the Interim Final Guidance recommends that the sample Coefficient of Variation be less than one; otherwise this "test" of Normality fails. A drawback to using the sample CV is that for Normally distributed

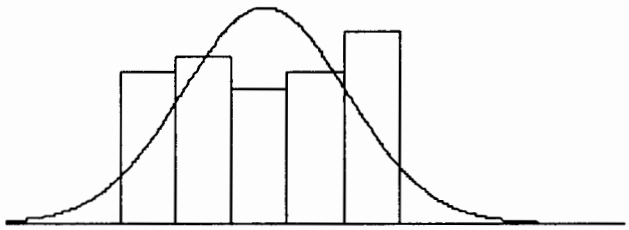
data, one can often get a sample CV greater than one when the true CV is only between 0.5 and 1. In other words, the sample CV, being a random variable, often estimates the true Coefficient of Variation with some error. Even if a Normal distribution model is appropriate, the Coefficient of Variation test may reject the model because the sample CV (but not the true CV) is too large.

The real purpose of the CV is to estimate the skewness of a dataset, not to test Normality. Truly Normal data can have any non-zero Coefficient of Variation, though the larger the CV, the greater the proportion of negative values predicted by the model. As such, a Normal distribution with large CV may be a poor model for positive concentration data. However, if the Coefficient of Variation test is used on the logarithms of the data to test Lognormality, negative logged concentrations will often be expected, nullifying the rationale used to support the CV test in the first place. A better way to estimate the skewness of a dataset is to compute the Coefficient of Skewness directly, as described below.

The Chi-square test is also recommended within the Interim Guidance. Though an acceptable goodness-of-fit test, it is not considered the most sensitive or powerful test of Normality in the current literature (Gan and Koehler, 1990). The major drawback to the Chi-square test can be explained by considering the behavior of parametric tests based on the Normal distribution. Most tests like the t-test or Analysis of Variance (ANOVA), which assume the underlying data to be Normally distributed, give fairly robust results when the Normality assumption fails over the middle ranges of the data distribution. That is, if the extreme tails are approximately Normal in shape even if the middle part of the density is not, these parametric tests will still tend to produce valid results. However, if the extreme tails are non-Normal in shape (e.g., highly skewed), Normal-based tests can lead to false conclusions, meaning that either a transformation of the data or a non-parametric technique should be used instead.

The Chi-square test entails a division of the sample data into bins or cells representing distinct, non-overlapping ranges of the data values (see figure below). In each bin, an expected value is computed based on the number of data points that would be found if the Normal distribution provided an appropriate model. The squared difference between the expected number and observed number is then computed and summed over all the bins to calculate the Chi-square test statistic.

CHI SQUARE GOODNESS OF FIT



If the Chi-square test indicates that the data are not Normally distributed, it may not be clear what ranges of the data most violate the Normality assumption. Departures from Normality in the middle bins are given nearly the same weight as departures from the extreme tail bins, and all the departures are summed together to form the test statistic. As such, the Chi-square test is not as powerful for detecting departures from Normality in the extreme tails of the data, the areas most crucial to the validity of parametric tests like the t-test or ANOVA (Miller, 1986). Furthermore, even if there are departures in the tails, but the middle portion of the data distribution is approximately Normal, the Chi-square test may not register as statistically significant in certain cases where better tests of Normality would. Because of this, four alternative, more sensitive tests of Normality are suggested below which can be used in conjunction with Probability Plots.

1.1.2 Probability Plots

As suggested within the Interim Final Guidance, a simple, yet useful graphical test for Normality is to plot the data on probability paper. The y-axis is scaled to represent probabilities according to the Normal distribution and the data are arranged in increasing order. An observed value is plotted on the x-axis and the proportion of observations less than or equal to each observed value is plotted as the y-coordinate. The scale is constructed so that, if the data are Normal, the points when plotted will approximate a straight line. Visually apparent curves or bends indicate that the data do not follow a Normal distribution (see Interim Final Guidance, pp. 4-8 to 4-11).

Probability Plots are particularly useful for spotting irregularities within the data when compared to a specific distributional model like the Normal. It is easy to determine whether departures from Normality are occurring more or less in the middle ranges of the data or in the extreme tails. Probability Plots can also indicate the presence of possible outlier values that do not follow the basic pattern of the data and can show the presence of significant positive or negative skewness.

If a (Normal) Probability Plot is done on the combined data from several wells and Normality is accepted, it implies that all of the data came from the same Normal distribution. Consequently, each subgroup of the data set (e.g., observations from distinct wells), has the same mean and standard deviation. If a Probability Plot is done on the data residuals (each value minus its subgroup mean) and is not a straight line, the interpretation is more complicated. In this case, either the residuals are not Normal, or there is a subgroup of the data with a Normal distribution but a different mean or standard deviation than the other subgroups. The Probability Plot will indicate a deviation from the underlying Normality assumption either way.

The same Probability Plot technique may be used to investigate whether a set of data or residuals follows the Lognormal distribution. The procedure is the same, except that one first replaces each observation by its natural logarithm. After the data have been transformed to their natural logarithms, the Probability Plot is constructed as before. The only difference is that the natural logarithms of the observations are used on the x-axis. If the data are Lognormal, the Probability Plot (on Normal probability paper) of the logarithms of the observations will approximate a straight line.

Many statistical software packages for personal computers will construct Probability Plots automatically with a simple command or two. If such software is available, there is no need to construct Probability Plots by hand or to obtain special graph paper. The plot itself may be generated somewhat differently than the method described above. In some packages, the observed value is plotted as before on the x-axis. The y-axis, however, now represents the quantile of the Normal distribution (often referred to as the "Normal score of the observation") corresponding to the cumulative probability of the observed value. The y-coordinate is often computed by the following formula:

$$y_i = \Phi^{-1}\left(\frac{i}{n+1}\right)$$

where Φ^{-1} denotes the inverse of the cumulative Normal distribution, n represents the sample size, and i represents the rank position of the ith ordered concentration. Since the computer does these calculations automatically, the formula does not have to be computed by hand.

EXAMPLE 1

Determine whether the following data set follows the Normal distribution by using a Probability Plot.

	Nickel Concentration (ppb)			
Month	Well 1	Well 2	Well 3	Well 4
1	58.8	19	39	3.1
2	1.0	81.5	151	942
3	262	331	27	85.6
4	56	14	21.4	10
5	8.7	64.4	578	637

SOLUTION

Step 1. List the measured nickel concentrations in order from lowest to highest.

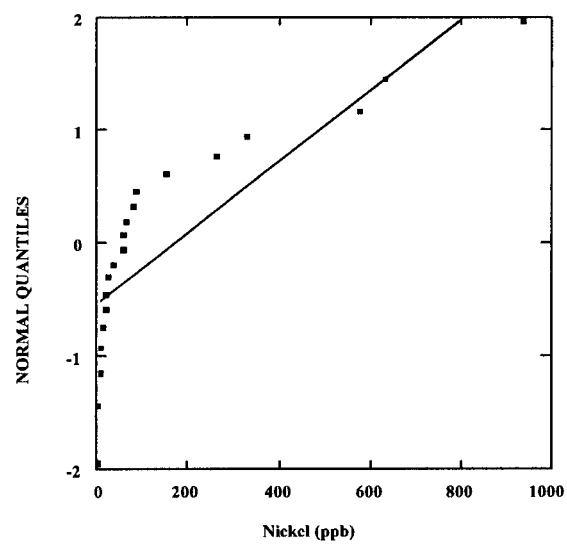
Nickel Concentration (ppb)	Order (i)	Probability 100*(i/(n+1))	Normal Quantile
1	1	5	-1.645
3.1	2	10	-1.28
8.7	3	14	-1.08
10	4	19	-0.88
14	5	24	-0.706
19	6	29	-0.55
21.4	7	33	-0.44
27	8	38	-0.305
39	9	43	-0.176
56	10	48	-0.05
58.8	11	52	0.05
64.4	12	57	0.176
81.5	13	62	0.305
85.6	14	67	0.44
151	15	71	0.55
262	16	76	0.706
331	17	81	0.88
578	18	86	1.08
637	19	90	1.28

942	20	95	1.645

Step 2. The cumulative probability is given in the third column and is computed as 100*(i/(n+1)) where n is the total number of samples (n=20). The last column gives the Normal quantiles corresponding to these probabilities.

Step 3. If using special graph paper, plot the probability versus the concentration for each sample. Otherwise, plot the Normal quantile versus the concentration for each sample, as in the plot below. The curvature found in the Probability Plot indicates that there is evidence of non-Normality in the data.

PROBABILITY PLOT



1.1.3 Coefficient of Skewness

The Coefficient of Skewness (γ_1) indicates to what degree a data set is skewed or asymmetric with respect to the mean. Data from a Normal distribution will have a Skewness Coefficient of zero, while asymmetric data will have a positive or negative skewness depending on whether the right- or left-hand tail of the distribution is longer and skinnier than the opposite tail.

Since ground-water monitoring concentration data are inherently nonnegative, one often expects the data to exhibit a certain degree of skewness. A small degree of skewness is not likely to affect the results of statistical tests based on an assumption of Normality. However, if the Skewness Coefficient is larger than 1 (in absolute value) and the sample size is small (e.g., n<25),

statistical research has shown that standard Normal theory-based tests are much less powerful than when the absolute skewness is less than 1 (Gayen, 1949).

Calculating the Skewness Coefficient is useful and not much more difficult than computing the Coefficient of Variation. It provides a quick indication of whether the skewness is minimal enough to assume that the data are roughly symmetric and hopefully Normal in distribution. If the original data exhibit a high Skewness Coefficient, the Normal distribution will provide a poor approximation to the data set. In that case, γ_1 can be computed on the logarithms of the data to test for symmetry of the logged data.

The Skewness Coefficient may be computed using the following formula:

$$\gamma_1 = \frac{\frac{1}{n}\sum_i (x_i - \bar{x})^3}{\left(\frac{n-1}{n}\right)^{\frac{3}{2}} (SD)^3}$$

where the numerator represents the average cubed residual and SD denotes the standard deviation of the measurements. Most statistics computer packages (e.g., Minitab, GEO-EAS) will compute the Skewness Coefficient automatically via a simple command.

EXAMPLE 2

Using the data in Example 1, compute the Skewness Coefficient to test for approximate symmetry in the data.

SOLUTION

Step 1. Compute the mean, standard deviation (SD), and average cubed residual for the nickel concentrations:

$$\bar{x} = 169.52 \text{ ppb}$$

$$SD = 259.72 \text{ ppb}$$

$$\frac{1}{n}\sum_i (x_i - \bar{x})^3 = 2.98923 * 10^8 \text{ ppb}^3$$

Step 2. Calculate the Coefficient of Skewness using the previous formula to get γ_1=1.84. Since the skewness is much larger than 1, the data appear to be significantly positively skewed. Do not assume that the data follow a Normal distribution.

Step 3. Since the original data evidence a high degree of skewness, one can attempt to compute the Skewness Coefficient on the logged data instead. In that case, the skewness works out to be $|\gamma_1| = 0.24 < 1$, indicating that the logged data values are slightly skewed, but not enough to reject an assumption of Normality in the logged data. In other words, the original data may be Lognormally distributed.

1.1.4 The Shapiro-Wilk Test of Normality ($n \leq 50$)

The Shapiro-Wilk test is recommended as a superior alternative to the Chi-square test for testing Normality of the data. It is based on the premise that if a set of data are Normally distributed, the ordered values should be highly correlated with corresponding quantiles taken from a Normal distribution (Shapiro and Wilk, 1965). In particular, the Shapiro-Wilk test gives substantial weight to evidence of non-Normality in the tails of a distribution, where the robustness of statistical tests based on the Normality assumption is most severely affected. The Chi-square test treats departures from Normality in the tails nearly the same as departures in the middle of a distribution, and so is less sensitive to the types of non-Normality that are most crucial. One cannot tell from a significant Chi-square goodness-of-fit test what sort of non-Normality is indicated.

The Shapiro-Wilk test statistic (W) will tend to be large when a Probability Plot of the data indicates a nearly straight line. Only when the plotted data show significant bends or curves will the test statistic be small. The Shapiro-Wilk test is considered to be one of the very best tests of Normality available (Miller, 1986; Madansky, 1988).

To calculate the test statistic W, one can use the following formula:

$$W = \left[\frac{b}{SD\sqrt{n-1}}\right]^2$$

where the numerator is computed as

$$b = \sum_{i=1}^{k} a_{n-i+1}\left(x_{(n-i+1)} - x_{(i)}\right) = \sum_{i=1}^{k} b_i$$

In this last formula, $x_{(j)}$ represents the jth smallest ordered value in the sample and coefficients a_j depend on the sample size n. The coefficients can be found for any sample size

from 3 up to 50 in Table A-1 of Appendix A. The value of k can be found as the greatest integer less than or equal to n/2.

Normality of the data should be rejected if the Shapiro-Wilk statistic is too low when compared to the critical values provided in Table A-2 of Appendix A. Otherwise one can assume the data are approximately Normal for purposes of further statistical analysis. As before, it is recommended that the test first be performed on the logarithms of the original data to test for Lognormality. If the logged data indicate non-Normality by the Shapiro-Wilk test, a re-test can be performed on the original data to test for Normality of the original concentrations.

EXAMPLE 3

Use the data of Example 1 to compute the Shapiro-Wilk test of Normality.

SOLUTION

Step 1. Order the data from smallest to largest and list, as in the following table. Also list the data in reverse order alongside the first column.

Step 2. Compute the differences $x_{(n-i+1)}-x_{(i)}$ in column 3 of the table by subtracting column 1 from column 2.

i	$x_{(i)}$	$x_{(n-i+1)}$	$x_{(n-i+1)}-x_{(i)}$	a_{n-i+1}	b_i
1	1.0	942.0	941.0	.4734	445.47
2	3.1	637.0	633.9	.3211	203.55
3	8.7	578.0	569.3	.2565	146.03
4	10.0	331.0	321.0	.2085	66.93
5	14.0	262.0	248.0	.1686	41.81
6	19.0	151.0	132.0	.1334	17.61
7	21.4	85.6	64.2	.1013	6.50
8	27.0	81.5	54.5	.0711	3.87
9	39.0	64.4	25.4	.0422	1.07
10	56.0	58.8	2.8	.0140	0.04
11	58.8	56.0	-2.8		b=932.88
12	64.4	39.0	-25.4		
13	81.5	27.0	-54.5		
14	85.6	21.4	-64.2		
15	151.0	19.0	-132.0		
16	262.0	14.0	-248.0		
17	331.0	10.0	-321.0		
18	578.0	8.7	-569.3		
19	637.0	3.1	-633.9		

20	942.0	1.0	-941.0

Step 3. Compute k as the greatest integer less than or equal to n/2. Since n=20, k=10 in this example.

Step 4. Look up the coefficients a_{n-i+1} from Table A-1 and list in column 4. Multiply the differences in column 3 by the coefficients in column 4 and add the first k products to get quantity b. In this case, b=932.88.

Step 5. Compute the standard deviation of the sample, SD=259.72. Then

$$W = \left[\frac{932.88}{259.72\sqrt{19}}\right]^2 = 0.679.$$

Step 6. Compare the computed value of W=0.679 to the 5% critical value for sample size 20 in Table A-2, namely $W_{.05,20}$=0.905. Since W < 0.905, the sample shows significant evidence of non-Normality by the Shapiro-Wilk test. The data should be transformed using natural logs and rechecked using the Shapiro-Wilk test before proceeding with further statistical analysis (Actually, the logged data should have been tested first. The original concentration data are used in this example to illustrate how the assumption of Normality can be rejected.)

1.1.5 The Shapiro-Francia Test of Normality (n>50)

The Shapiro-Wilk test of Normality can be used for sample sizes up to 50. When the sample is larger than 50, a slight modification of the procedure called the Shapiro-Francia test (Shapiro and Francia, 1972) can be used instead.

Like the Shapiro-Wilk test, the Shapiro-Francia test statistic (W′) will tend to be large when a Probability Plot of the data indicates a nearly straight line. Only when the plotted data show significant bends or curves will the test statistic be small.

To calculate the test statistic W′, one can use the following formula:

$$W' = \frac{\left[\Sigma_i m_i x_{(i)}\right]^2}{(n-1)SD^2 \Sigma_i m_i^2}$$

where $x_{(i)}$ represents the ith ordered value of the sample and where m_i denotes the approximate expected value of the ith ordered Normal quantile. The values for m_i can be approximately computed as

$$m_i = \Phi^{-1}\left(\frac{i}{n+1}\right)$$

where Φ^{-1} denotes the inverse of the standard Normal distribution with zero mean and unit variance. These values can be computed by hand using a Normal probability table or via simple commands in many statistical computer packages.

Normality of the data should be rejected if the Shapiro-Francia statistic is too low when compared to the critical values provided in Table A-3 of Appendix A. Otherwise one can assume the data are approximately Normal for purposes of further statistical analysis. As before, the logged data should be tested first to see if a Lognormal model is appropriate. If these data indicate non-Normality by the Shapiro-Francia test, a re-test can be performed on the original data.

1.1.6 The Probability Plot Correlation Coefficient

One other alternative test for Normality that is roughly equivalent to the Shapiro-Wilk and Shapiro-Francia tests is the Probability Plot Correlation Coefficient test described by Filliben (1975). This test fits in perfectly with the use of Probability Plots, because the essence of the test is to compute the common correlation coefficient for points on a Probability Plot. Since the correlation coefficient is a measure of the linearity of the points on a scatterplot, the Probability Plot Correlation Coefficient, like the Shapiro-Wilk test, will be high when the plotted points fall along a straight line and low when there are significant bends and curves in the Probability Plot. Comparison of the Shapiro-Wilk and Probability Plot Correlation Coefficient tests has indicated very similar statistical power for detecting non-Normality (Ryan and Joiner, 1976).

The construction of the test statistic is somewhat different from the Shapiro-Wilk W, but not difficult to implement. Also, tabled critical values for the correlation coefficient have been derived for sample sizes up to n=100 (and are reproduced in Table A-4 of Appendix A). The Probability Plot Correlation Coefficient may be computed as

$$r = \frac{\sum_{i=1}^{n} X_{(i)} M_i - n\overline{X}\overline{M}}{C_n \times SD\sqrt{n-1}}$$

where $X_{(i)}$ represents the ith smallest ordered concentration value, M_i is the median of the ith order statistic from a standard Normal distribution, and $\overline{X}$ and $\overline{M}$ represent the average values of $X_{(i)}$ and $M_{(i)}$. The ith Normal order statistic median may be approximated as $M_i = \Phi^{-1}(m_i)$, where as before, Φ^{-1} is the inverse of the standard Normal cumulative distribution and m_i can be computed as follows (given sample size n):

$$m_i = \begin{cases} 1-(.5)^{1/n} & \text{for } i=1 \\ (i-.3175)/(n+.365) & \text{for } 1<i<n \\ (.5)^{1/n} & \text{for } i=n \end{cases}$$

Quantity C_n represents the square root of the sum of squares of the M_i's minus n times the average value $\overline{M}$, that is

$$C_n = \sqrt{\sum_i M_i^2 - n\overline{M}^2}$$

When working with a complete sample (i.e., containing no nondetects or censored values), the average value $\overline{M}=0$, and so the formula for the Probability Plot Correlation Coefficient simplifies to

$$\boxed{r = \frac{\sum_i X_{(i)} M_i}{\sqrt{\sum_i M_i^2} \times SD\sqrt{n-1}}}$$

EXAMPLE 4

Use the data of Example 1 to compute the Probability Plot Correlation Coefficient test.

SOLUTION

Step 1. Order the data from smallest to largest and list, as in the following table.

Step 2. Compute the quantities m_i from Filliben's formula above for each i in column 2 and the order statistic medians, M_i, in column 3 by applying the inverse Normal transformation to column 2.

Step 3. Since this sample contains no nondetects, the simplified formula for r may be used. Compute the products $X_{(i)}*M_i$ in column 4 and sum to get the numerator of the

correlation coefficient (equal to 3,836.81 in this case). Also compute M_i^2 in column 5 and sum to find quantity C_n^2=17.12.

i	$x_{(i)}$	m_i	M_i	$X_{(i)}*M_i$	M_i^2
1	1.0	.03406	-1.8242	-1.824	3.328
2	3.1	.08262	-1.3877	-4.302	1.926
3	8.7	.13172	-1.1183	-9.729	1.251
4	10.0	.18082	-0.9122	-9.122	0.832
5	14.0	.22993	-0.7391	-10.347	0.546
6	19.0	.27903	-0.5857	-11.129	0.343
7	21.4	.32814	-0.4451	-9.524	0.198
8	27.0	.37724	-0.3127	-8.444	0.098
9	39.0	.42634	-0.1857	-7.242	0.034
10	56.0	.47545	-0.0616	-3.448	0.004
11	58.8	.52455	0.0616	3.621	0.004
12	64.4	.57366	0.1857	11.959	0.034
13	81.5	.62276	0.3127	25.488	0.098
14	85.6	.67186	0.4451	38.097	0.198
15	151.0	.72097	0.5857	88.445	0.343
16	262.0	.77007	0.7391	193.638	0.546
17	331.0	.81918	0.9122	301.953	0.832
18	578.0	.86828	1.1183	646.376	1.251
19	637.0	.91738	1.3877	883.941	1.926
20	942.0	.96594	1.8242	1718.408	3.328

Step 4. Compute the Probability Plot Correlation Coefficient using the simplified formula for r, where SD=259.72 and C_n=4.1375, to get

$$r = \frac{3836.81}{(4.1375)(259.72)\sqrt{19}} = 0.819$$

Step 5. Compare the computed value of r=0.819 to the 5% critical value for sample size 20 in Table A-4, namely $R_{.05,20}$=0.950. Since r < 0.950, the sample shows significant evidence of non-Normality by the Probability Plot Correlation Coefficient test. The data should be transformed using natural logs and the correlation coefficient recalculated before proceeding with further statistical analysis.

EXAMPLE 5

The data in Examples 1, 2, 3, and 4 showed significant evidence of non-Normality. Instead of first logging the concentrations before testing for Normality, the original data were used. This was done to illustrate why the Lognormal distribution is usually a better default model than the Normal. In this example, use the same data to determine whether the measurements better follow a Lognormal distribution.

Computing the natural logarithms of the data gives the table below.

	Logged Nickel Concentrations log (ppb)			
Month	Well 1	Well 2	Well 3	Well 4
1	4.07	2.94	3.66	1.13
2	0.00	4.40	5.02	6.85
3	5.57	5.80	3.30	4.45
4	4.03	2.64	3.06	2.30
5	2.16	4.17	6.36	6.46

SOLUTION

Method 1. Probability Plots

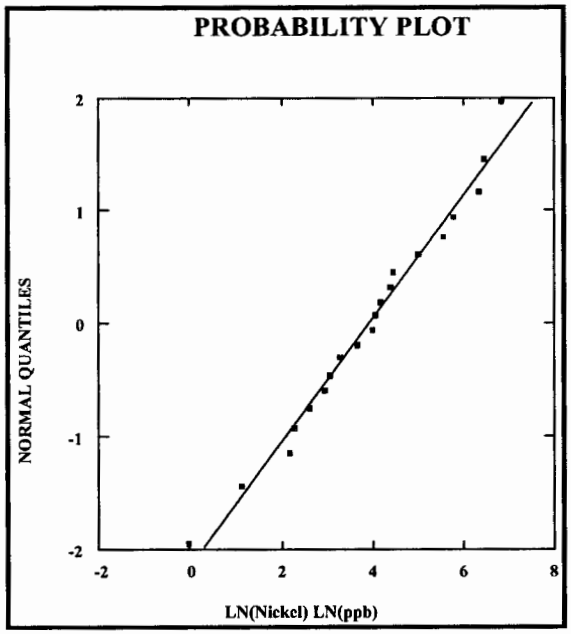
Step 1. List the natural logarithms of the measured nickel concentrations in order from lowest to highest.

Order (i)	Log Nickel Concentration log(ppb)	Probability 100*(i/(n+1))	Normal Quantiles
1	0.00	5	-1.645
2	1.13	10	-1.28
3	2.16	14	-1.08
4	2.30	19	-0.88
5	2.64	24	-0.706
6	2.94	29	-0.55
7	3.06	33	-0.44
8	3.30	38	-0.305
9	3.66	43	-0.176
10	4.03	48	-0.05
11	4.07	52	0.05

12	4.17	57	0.176
13	4.40	62	0.305
14	4.45	67	0.44
15	5.02	71	0.55
16	5.57	76	0.706
17	5.80	81	0.88
18	6.36	86	1.08
19	6.46	90	1.28
20	6.85	95	1.645

Step 2. Compute the probability as shown in the third column by calculating 100*(i/n+1), where n is the total number of samples (n=20). The corresponding Normal quantiles are given in column 4.

Step 3. Plot the Normal quantiles against the natural logarithms of the observed concentrations to get the following graph. The plot indicates a nearly straight line fit (verified by calculation of the Correlation Coefficient given in Method 4). There is no substantial evidence that the data do not follow a Lognormal distribution. The Normal-theory procedure(s) should be performed on the log-transformed data.



Method 2. Coefficient of Skewness

Step 1. Calculate the mean, SD, and average cubed residuals of the natural logarithms of the data.

$$\bar{x} = 3.918 \log(\text{ppb})$$
$$SD = 1.802 \log(\text{ppb})$$

$$\frac{1}{n}\Sigma_i(x_i - \bar{x})^3 = -1.325 \log^3(\text{ppb})$$

Step 2. Calculate the Skewness Coefficient, γ_1.

$$\gamma_1 = \frac{-1.325}{(.95)^{\frac{3}{2}}(1.802)^3} = -0.244$$

Step 3. Compute the absolute value of the skewness, $|\gamma_1|=|-0.244|=0.244$.

Step 4. Since the absolute value of the Skewness Coefficient is less than 1, the data do not show evidence of significant skewness. A Normal approximation to the log-transformed data may therefore be appropriate, but this model should be further checked.

Method 3. Shapiro-Wilk Test

Step 1. Order the logged data from smallest to largest and list, as in following table. Also list the data in reverse order and compute the differences $x_{(n-i+1)}-x_{(i)}$.

i	$LN(x_{(i)})$	$LN(x_{(n-i+1)})$	a_{n-i+1}	b_i
1	0.00	6.85	.4734	3.24
2	1.13	6.46	.3211	1.71
3	2.16	6.36	.2565	1.08
4	2.30	5.80	.2085	0.73
5	2.64	5.57	.1686	0.49
6	2.94	5.02	.1334	0.28
7	3.06	4.45	.1013	0.14
8	3.30	4.40	.0711	0.08
9	3.66	4.17	.0422	0.02
10	4.03	4.07	.0140	0.00
11	4.07	4.03		b=7.77
12	4.17	3.66		
13	4.40	3.30		
14	4.45	3.06		
15	5.02	2.94		

16	5.57	2.64
17	5.80	2.30
18	6.36	2.16
19	6.46	1.13
20	6.85	0.00

Step 2. Compute k=10, since n/2=10. Look up the coefficients a_{n-i+1} from Table A-1 and multiply by the first k differences between columns 2 and 1 to get the quantities b_i. Add these 10 products to get b=7.77.

Step 3. Compute the standard deviation of the logged data, SD=1.8014. Then the Shapiro-Wilk statistic is given by

$$W = \left[\frac{7.77}{1.8014\sqrt{19}}\right]^2 = 0.979.$$

Step 4. Compare the computed value of W to the 5% critical value for sample size 20 in Table A-2, namely $W_{.05,20}$=0.905. Since W=0.979>0.905, the sample shows no significant evidence of non-Normality by the Shapiro-Wilk test. Proceed with further statistical analysis using the log-transformed data.

Method 4. Probability Plot Correlation Coefficient

Step 1. Order the logged data from smallest to largest and list below.

Order (i)	Log Nickel Concentration log(ppb)	m_i	M_i	$X_{(i)}*M_i$	M_i^2
1	0.00	.03406	-1.8242	0.000	3.328
2	1.13	.08262	-1.3877	-1.568	1.926
3	2.16	.13172	-1.1183	-2.416	1.251
4	2.30	.18082	-0.9122	-2.098	0.832
5	2.64	.22993	-0.7391	-1.951	0.546
6	2.94	.27903	-0.5857	-1.722	0.343
7	3.06	.32814	-0.4451	-1.362	0.198
8	3.30	.37724	-0.3127	-1.032	0.098
9	3.66	.42634	-0.1857	-0.680	0.034
10	4.03	.47545	-0.0616	-0.248	0.004
11	4.07	.52455	0.0616	0.251	0.004
12	4.17	.57366	0.1857	0.774	0.034
13	4.40	.62276	0.3127	1.376	0.098
14	4.45	.67186	0.4451	1.981	0.198

15	5.02	.72097	0.5857	2.940	0.343
16	5.57	.77007	0.7391	4.117	0.546
17	5.80	.81918	0.9122	5.291	0.832
18	6.36	.86828	1.1183	7.112	1.251
19	6.46	.91738	1.3877	8.965	1.926
20	6.85	.96594	1.8242	12.496	3.328

Step 2. Compute the quantities m_i and the order statistic medians M_i, according to the procedure in Example 4 (note that these values depend only on the sample size and are identical to the quantities in Example 4).

Step 3. Compute the products $X_{(i)}*M_i$ in column 4 and sum to get the numerator of the correlation coefficient (equal to 32.226 in this case). Also compute M_i^2 in column 5 and sum to find quantity $C_n^2=17.12$.

Step 4. Compute the Probability Plot Correlation Coefficient using the simplified formula for r, where SD=1.8025 and C_n=4.1375, to get

$$r = \frac{32.226}{(4.1375)(1.8025)\sqrt{19}} = 0.991$$

Step 5. Compare the computed value of r=0.991 to the 5% critical value for sample size 20 in Table A-4, namely $R_{.05,20}$=0.950. Since r > 0.950, the logged data show no significant evidence of non-Normality by the Probability Plot Correlation Coefficient test. Therefore, Lognormality of the original data could be assumed in subsequent statistical procedures.

1.2 TESTING FOR HOMOGENEITY OF VARIANCE

One of the most important assumptions for the parametric analysis of variance (ANOVA) is that the different groups (e.g., different wells) have approximately the same variance. If this is not the case, the power of the F-test (its ability to detect differences among the group means) is reduced. Mild differences in variance are not too bad. The effect becomes noticeable when the largest and smallest group variances differ by a ratio of about 4 and becomes quite severe when the ratio is 10 or more (Milliken and Johnson, 1984).

The procedure suggested in the EPA guidance document, Bartlett's test, is one way to test whether the sample data give evidence that the well groups have different variances. However, Bartlett's test is sensitive to non-Normality in the data and may give misleading results unless one knows in advance that the data are approximately Normal (Milliken and Johnson, 1984). As an

alternative to Bartlett's test, two procedures for testing homogeneity of the variances are described below that are less sensitive to non-Normality.

1.2.1 Box Plots

Box Plots were first developed for exploratory data analysis as a quick way to visualize the "spread" or dispersion within a data set. In the context of variance testing, one can construct a Box Plot for each well group and compare the boxes to see if the assumption of equal variances is reasonable. Such a comparison is not a formal test procedure, but is easier to perform and is often sufficient for checking the group variance assumption.

The idea behind a Box Plot is to order the data from lowest to highest and to trim off 25 percent of the observations on either end, leaving just the middle 50 percent of the sample values. The spread between the lowest and highest values of this middle 50 percent (known as the interquartile range or IQR) is represented by the length of the box. The very middle observation (i.e., the median) can also be shown as a line cutting the box in two.

To construct a Box Plot, calculate the median and upper and lower quantiles of the data set (respectively, the 50th, 25th, and 75th percentiles). To do this, calculate $k=p(n+1)/100$ where n=number of samples and p=percentile of interest. If k is an integer, let the kth ordered or ranked value be an estimate of the pth percentile of the data. If k is not an integer, let the pth percentile be equal to the average of the two values closest in rank position to k. For example, if the data set consists of the 10 values {1, 4, 6.2, 10, 15, 17.1, 18, 22, 25, 30.5}, the position of the median would be found as $50*(10+1)/100=5.5$. The median would then be computed as the average of the 5th and 6th ordered values, or $(15+17.1)/2=16.05$.

Likewise, the position of the lower quartile would be $25*(10+1)/100=2.75$. Calculate the average of the 2nd and 3rd ordered observations to estimate this percentile, i.e., $(4+6.2)/2=5.1$. Since the upper quartile is found to be 23.5, the length of Box Plot would be the difference between the upper and lower quartiles, or $(23.5-5.1)=18.4$. The box itself should be drawn on a graph with the y-axis representing concentration and the x-axis denoting the wells being plotted. Three horizontal lines are drawn for each well, one line each at the lower and upper quartiles and another at the median concentration. Vertical connecting lines are drawn to complete the box.

Most statistics packages can directly calculate the statistics needed to draw a Box Plot, and many will construct the Box Plots as well. In some computer packages, the Box Plot will also

have two "whiskers" extending from the edges of the box. These lines indicate the positions of extreme values in the data set, but generally should not be used to approximate the overall dispersion.

If the box length for each group is less than 3 times the length of the shortest box, the sample variances are probably close enough to assume equal group variances. If, however, the box length for any group is at least triple the length of the box for another group, the variances may be significantly different (Kirk Cameron, SAIC, personal communication). In that case, the data should be further checked using Levene's test described in the following section. If Levene's test is significant, the data may need to be transformed or a non-parametric rank procedure considered before proceeding with further analysis.

EXAMPLE 6

Construct Box Plots for each well group to test for equality of variances.

	Arsenic Concentration (ppm)					
Month	Well 1	Well 2	Well 3	Well 4	Well 5	Well 6
1	22.9	2.0	2.0	7.84	24.9	0.34
2	3.09	1.25	109.4	9.3	1.3	4.78
3	35.7	7.8	4.5	25.9	0.75	2.85
4	4.18	52	2.5	2.0	27	1.2

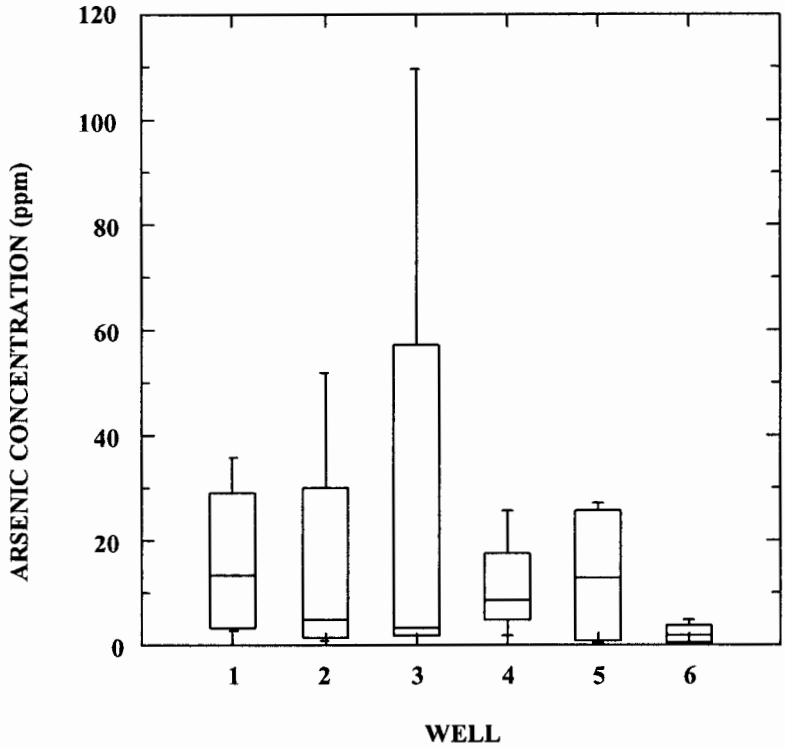
SOLUTION

Step 1. Compute the 25th, 50th, and 75th percentiles for the data in each well group. To calculate the pth percentile by hand, order the data from lowest to highest. Calculate p*(n+1)/100 to find the ordered position of the pth percentile. If necessary, interpolate between sample values to estimate the desired percentile.

Step 2. Using well 1 as an example, n+1=5 (since there are 4 data values). To calculate the 25th percentile, compute its ordered position (i.e., rank) as 25*5/100=1.25. Average the 1st and 2nd ranked values at well 1 (i.e., 3.09 and 4.18) to find an estimated lower quartile of 3.64. This estimate gives the lower end of the Box Plot. The upper end or 75th percentile can be computed similarly as the average of the 3rd and 4th ranked values, or (22.9+35.7)/2=29.3. The median is the average of the 2nd and 3rd ranked values, giving an estimate of 13.14.

Step 3. Construct Box Plots for each well group, lined up side by side on the same axes.

BOX PLOTS OF WELL DATA



Step 4. Since the box length for well 3 is more than three times the box lengths for wells 4 and 6, there is evidence that the group variances may be significantly different. These data should be further checked using Levene's test described in the next section.

1.2.2 Levene's Test

Levene's test is a more formal procedure than Box Plots for testing homogeneity of variance that, unlike Bartlett's test, is not sensitive to non-Normality in the data. Levene's test has been shown to have power nearly as great as Bartlett's test for Normally distributed data and power superior to Bartlett's for non-Normal data (Milliken and Johnson, 1984).

To conduct Levene's test, first compute the new variables

$$z_{ij} = \left| x_{ij} - \bar{x}_i \right|$$

where x_{ij} represents the jth value from the ith well and $\bar{x}_i$ is the ith well mean. The values z_{ij} represent the absolute values of the usual residuals. Then run a standard one-way analysis of variance (ANOVA) on the variables z_{ij}. If the F-test is significant, reject the hypothesis of equal group variances. Otherwise, proceed with analysis of the x_{ij}'s as initially planned.

EXAMPLE 7

Use the data from Example 6 to conduct Levene's test of equal variances.

SOLUTION

Step 1. Calculate the group mean for each well $(\bar{x}_i)$

Well 1 mean = 16.47 Well 4 mean = 11.26

Well 2 mean = 15.76 Well 5 mean = 13.49

Well 3 mean = 29.60 Well 6 mean = 2.29

Step 2. Compute the absolute residuals z_{ij} in each well and the well means of the residuals $(\bar{z}_i)$.

	Absolute Residuals					
Month	Well 1	Well 2	Well 3	Well 4	Well 5	Well 6
1	6.43	13.76	27.6	3.42	11.41	1.95
2	13.38	14.51	79.8	1.96	12.19	2.49
3	19.23	7.96	25.1	14.64	12.74	0.56
4	12.29	36.24	27.1	9.26	13.51	1.09
Well Mean $(\bar{z}_i)$	= 12.83	18.12	39.9	7.32	12.46	1.52
Overall Mean $(\bar{z})$	= 15.36					

Step 3. Compute the sums of squares for the absolute residuals.

$$SS_{TOTAL} = (N-1)\ SDz^2 = 6300.89$$

$$SS_{WELLS} = \sum_i n_i \bar{z}_i^2 - N\bar{z}^2 = 3522.90$$

$$SS_{ERROR} = SS_{TOTAL} - SS_{WELLS} = 2777.99$$

Step 4. Construct an analysis of variance table to calculate the F-statistic. The degrees of freedom (df) are computed as (#groups–1)=(6–1)=5 df and (#samples–#groups)=(24–6)=18 df.

ANOVA Table

Source	Sum-of-Squares	df	Mean-Square	F-Ratio	P
Between Wells	3522.90	5	704.58	4.56	0.007
Error	2777.99	18	154.33		
Total	6300.89	23			

Step 5. Since the F-statistic of 4.56 exceeds the tabulated value of $F_{.05}$=2.77 with 5 and 18 df, the assumption of equal variances should be rejected. Since the original concentration data are used in this example, the data should be logged and retested.

2. RECOMMENDATIONS FOR HANDLING NONDETECTS

The basic recommendations within the Interim Final Guidance for handling nondetect analyses include the following (see p. 8-2): 1) if less than 15 percent of all samples are nondetect, replace each nondetect by half its detection or quantitation limit and proceed with a parametric analysis, such as ANOVA, Tolerance Limits, or Prediction Limits; 2) if the percent of nondetects is between 15 and 50, either use Cohen's adjustment to the sample mean and variance in order to proceed with a parametric analysis, or employ a non-parametric procedure by using the ranks of the observations and by treating all nondetects as tied values; 3) if the percent of nondetects is greater than 50 percent, use the Test of Proportions.

As to the first recommendation, experience at EPA and research at the United States Geological Survey (USGS, Dennis Helsel, personal communication, 1991) has indicated that if less than 15 percent of the samples are nondetect, the results of parametric statistical tests will not be substantially affected if nondetects are replaced by half their detection limits. When more than 15 percent of the samples are nondetect, however, the handling of nondetects is more crucial to the outcome of statistical procedures. Indeed, simple substitution methods tend to perform poorly in statistical tests when the nondetect percentage is substantial (Gilliom and Helsel, 1986).

Even with a small proportion of nondetects, however, care should be taken when choosing between the method detection limit (MDL) and the practical quantitation limit (PQL) in characterizing "nondetect" concentrations. Many nondetects are characterized by analytical laboratories with one of three data qualifier flags: "U," "J," or "E." Samples with a "U" data qualifier represent "undetected" measurements, meaning that the signal characteristic of that analyte could not be observed or distinguished from "background noise" during lab analysis. Inorganic samples with an "E" flag and organic samples with a "J" flag may or may not be reported with an estimated concentration. If no concentration is estimated, these samples represent "detected but not quantified" measurements. In this case, the actual concentration is assumed to be positive, but somewhere between zero and the PQL. Since all of these non-detects may or may not have actual positive concentrations between zero and the PQL, the suggested substitution for parametric statistical procedures is to replace each nondetect by one-half the PQL (note, however, that "E" and "J" samples reported with estimated concentrations should be treated, for statistical purposes, as valid measurements. Substitution of one-half the PQL is <u>not recommended</u> for these samples).

In no case should nondetect concentrations be assumed to be bounded above by the MDL. The MDL is estimated on the basis of ideal laboratory conditions with ideal analyte samples and does not account for matrix or other interferences encountered when analyzing specific, actual field samples. For this reason, the PQL should be taken as the most reasonable upper bound for nondetect concentrations.

It should also be noted that the distinction between "undetected" and "detected but not quantified" measurements has more specific implications for rank-based non-parametric procedures. Rather than assigning the same tied rank to all nondetects (see below and in **Section 3**), "detected but not quantified" measurements should be given larger ranks than those assigned to "undetected" samples. In fact the two types of nondetects should be treated as two distinct groups of tied observations for use in the Wilcoxon and Kruskal-Wallis non-parametric procedures.

2.1 NONDETECTS IN ANOVA PROCEDURES

For a moderate to large percentage of nondetects (i.e., over 15%), the handling of nondetects should vary depending on the statistical procedure to be run. If background data from one or more upgradient wells are to be compared simultaneously with samples from one or more downgradient wells via a t-test or ANOVA type procedure, the simplest and most reliable recommendation is to switch to a non-parametric analysis. The distributional assumptions for parametric procedures can be rather difficult to check when a substantial fraction of nondetects exists. Furthermore, the non-parametric alternatives described in **Section 3** tend to be efficient at detecting contamination when the underlying data are Normally distributed, and are often more powerful than the parametric methods when the underlying data do not follow a Normal distribution.

Nondetects are handled easily in a nonparametric analysis. All data values are first ordered and replaced by their ranks. Nondetects are treated as tied values and replaced by their midranks (see **Section 3**). Then a Wilcoxon Rank-Sum or Kruskal-Wallis test is run on the ranked data depending on whether one or more than one downgradient well is being tested.

The Test of Proportions is not recommended in this Addendum, even if the percentage of nondetects is over 50 percent. Instead, for all two-group comparisons that involve more than 15 percent nondetects, the non-parametric Wilcoxon Rank-Sum procedure is recommended. Although acceptable as a statistical procedure, the Test of Proportions does not account for

potentially different magnitudes among the concentrations of detected values. Rather, each sample is treated as a 0 or 1 depending on whether the measured concentration is below or above the detection limit. The Test of Proportions ignores information about concentration magnitudes, and hence is usually less powerful than a non-parametric rank-based test like the Wilcoxon Rank-Sum, even after adjusting for a large fraction of tied observations (e.g., nondetects). This is because the ranks of a dataset preserve additional information about the relative magnitudes of the concentration values, information which is lost when all observations are scored as 0's and 1's.

Another drawback to the Test of Proportions, as presented in the Interim Final Guidance, is that the procedure relies on a Normal probability approximation to the Binomial distribution of 0's and 1's. This approximation is recommended only when the quantities $n \times$ (%NDs) and $n \times (1-$%NDs) are no smaller than 5. If the percentage of nondetects is quite high and/or the sample size is fairly small, these conditions may be violated, leading potentially to inaccurate results.

Comparison of the Test of Proportions to the Wilcoxon Rank-Sum test shows that for small to moderate proportions of nondetects (say 0 to 60 percent), the Wilcoxon Rank-Sum procedure adjusted for ties is more powerful in identifying real concentration differences than the Test of Proportions. When the percentage of nondetects is quite high (at least 70 to 75 percent), the Test of Proportions appears to be slightly more powerful in some cases than the Wilcoxon, but the results of the two tests almost always lead to the same conclusion, so it makes sense to simply recommend the Wilcoxon Rank-Sum test in all cases where nondetects constitute more than 15 percent of the samples.

2.2 NONDETECTS IN STATISTICAL INTERVALS

If the chosen method is a statistical interval (Confidence, Tolerance or Prediction limit) used to compare background data against each downgradient well separately, more options are available for handling moderate proportions of nondetects. The basis of any parametric statistical interval limit is the formula $\bar{x} \pm \kappa \cdot s$, where $\bar{x}$ and s represent the sample mean and standard deviation of the (background) data and κ depends on the interval type and characteristics of the monitoring network. To use a parametric interval in the presence of a substantial number of nondetects, it is necessary to estimate the sample mean and standard deviation. But since nondetect concentrations are unknown, simple formulas for the mean and standard deviation cannot be computed directly. Two basic approaches to estimating or "adjusting" the mean and standard deviation in this situation have been described by Cohen (1959) and Aitchison (1955).

The underlying assumptions of these procedures are somewhat different. Cohen's adjustment (which is described in detail on pp. 8-7 to 8-11 of the Interim Final Guidance) assumes that all the data (detects and nondetects) come from the same Normal or Lognormal population, but that nondetect values have been "censored" at their detection limits. This implies that the contaminant of concern is present in nondetect samples, but the analytical equipment is not sensitive to concentrations lower than the detection limit. Aitchison's adjustment, on the other hand, is constructed on the assumption that nondetect samples are free of contamination, so that all nondetects may be regarded as zero concentrations. In some situations, particularly when the analyte of concern has been detected infrequently in background measurements, this assumption may be practical, even if it cannot be verified directly.

Before choosing between Cohen's and Aitchison's approaches, it should be cautioned that Cohen's adjustment may not give valid results if the proportion of nondetects exceeds 50%. In a case study by McNichols and Davis (1988), the false positive rate associated with the use of t-tests based on Cohen's method rose substantially when the fraction of nondetects was greater than 50%. This occurred because the adjusted estimates of the mean and standard deviation are more highly correlated as the percentage of nondetects increases, leading to less reliable statistical tests (including statistical interval tests).

On the other hand, with less than 50% nondetects, Cohen's method performed adequately in the McNichols and Davis case study, provided the data were not overly skewed and that more extensive tables than those included within the Interim Final Guidance were available to calculate Cohen's adjustment parameter. As a remedy to the latter caveat, a more extensive table of Cohen's adjustment parameter is provided in Appendix A (Table A-5). It is also recommended that the data (detected measurements and nondetect detection limits) first be log-transformed prior to computing either Cohen's or Aitchison's adjustment, especially since both procedures assume that the underlying data are Normally distributed.

2.2.1 Censored and Detects-Only Probability Plots

To decide which approach is more appropriate for a particular set of ground water data, two separate Probability Plots can be constructed. The first is called a Censored Probability Plot and is a test of Cohen's underlying assumption. In this method, the combined set of detects and nondetects is ordered (with nondetects being given arbitrary but distinct ranks). Cumulative probabilities or Normal quantiles (see **Section 1.1**) are then computed for the data set as in a regular Probability Plot. However, only the detected values and their associated Normal quantiles

are actually plotted. If the shape of the Censored Probability Plot is reasonably linear, then Cohen's assumption that nondetects have been "censored" at their detection limit is probably acceptable and Cohen's adjustment can be made to estimate the sample mean and standard deviation. If the Censored Probability Plot has significant bends and curves, particularly in one or both tails, one might consider Aitchison's procedure instead.

To test the assumptions of Aitchison's method, a Detects-Only Probability Plot may be constructed. In this case, nondetects are completely ignored and a standard Probability Plot is constructed using only the detected measurements. Thus, cumulative probabilities or Normal quantiles are computed only for the ordered detected values. Comparison of a Detects-Only Probability Plot with a Censored Probability Plot will indicate that the same number of points and concentration values are plotted on each graph. However, different Normal quantiles are associated with each detected concentration. If the Detects-Only Probability Plot is reasonably linear, then the assumptions underlying Aitchison's adjustment (i.e., that "nondetects" represent zero concentrations, and that detects and nondetects follow separate probability distributions) are probably reasonable.

If it is not clear which of the Censored or Detects-Only Probability Plots is more linear, Probability Plot Correlation Coefficients can be computed for both approaches (note that the correlations should only involve the points actually plotted, that is, detected concentrations). The plot with the higher correlation coefficient will represent the most linear trend. Be careful, however, to use other, non-statistical judgments to help decide which of Cohen's and Aitchison's underlying assumptions appears to be most reasonable based on the specific characteristics of the data set. It is also likely that these Probability Plots may have to be constructed on the logarithms of the data instead of the original values, if in fact the most appropriate underlying distribution is the Lognormal instead of the Normal.

EXAMPLE 8

Create Censored and Detects-Only Probability Plots with the following zinc data to determine whether Cohen's adjustment or Aitchison's adjustment is most appropriate for estimating the true mean and standard deviation.

	Zinc Concentrations (ppb) at Background Wells				
Sample	Well 1	Well 2	Well 3	Well 4	Well 5
1	<7	<7	<7	11.69	<7
2	11.41	<7	12.85	10.90	<7
3	<7	13.70	14.20	<7	<7
4	<7	11.56	9.36	12.22	11.15
5	<7	<7	<7	11.05	13.31
6	10.00	<7	12.00	<7	12.35
7	15.00	10.50	<7	13.24	<7
8	<7	12.59	<7	<7	8.74

SOLUTION

Step 1. Pool together the data from the five background wells and list in order in the table below.

Step 2. To construct the Censored Probability Plot, compute the probabilities i/(n+1) using the combined set of detects and nondetects, as in column 3. Find the Normal quantiles associated with these probabilities by applying the inverse standard Normal transformation, Φ^{-1}.

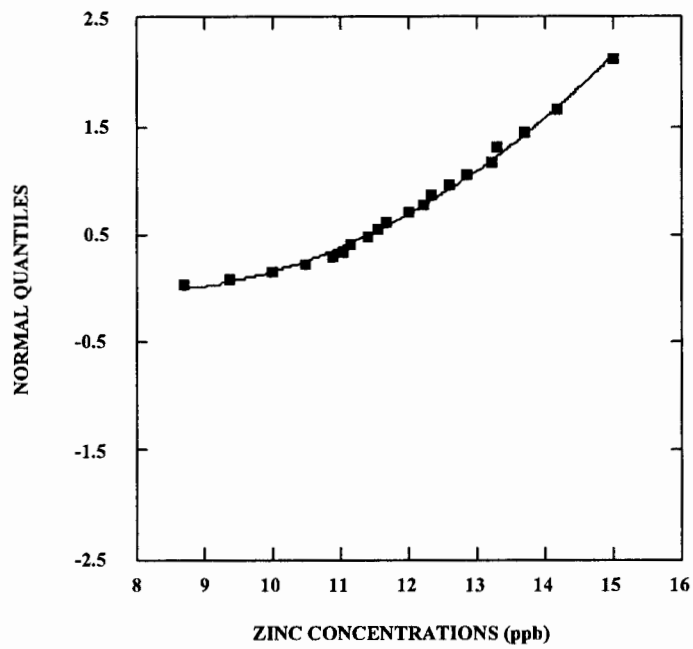
Step 3. To construct the Detects-Only Probability Plot, compute the probabilities in column 5 using only the detected zinc values. Again apply the inverse standard Normal transformation to find the associated Normal quantiles in column 6. Note that nondetects are ignored completely in this method.

Order (i)	Zinc Conc. (ppb)	Censored Probs.	Normal Quantiles	Detects-Only Probs.	Normal Quantiles
1	<7	.024	-1.971		
2	<7	.049	-1.657		
3	<7	.073	-1.453		
4	<7	.098	-1.296		
5	<7	.122	-1.165		
6	<7	.146	-1.052		
7	<7	.171	-0.951		
8	<7	.195	-0.859		
9	<7	.220	-0.774		
10	<7	.244	-0.694		
11	<7	.268	-0.618		
12	<7	.293	-0.546		
13	<7	.317	-0.476		
14	<7	.341	-0.408		
15	<7	.366	-0.343		
16	<7	.390	-0.279		
17	<7	.415	-0.216		
18	<7	.439	-0.153		
19	<7	.463	-0.092		
20	<7	.488	-0.031		
21	8.74	.512	0.031	.048	-1.668
22	9.36	.537	0.092	.095	-1.309
23	10.00	.561	0.153	.143	-1.068
24	10.50	.585	0.216	.190	-0.876
25	10.90	.610	0.279	.238	-0.712
26	11.05	.634	0.343	.286	-0.566
27	11.15	.659	0.408	.333	-0.431
28	11.41	.683	0.476	.381	-0.303
29	11.56	.707	0.546	.429	-0.180
30	11.69	.732	0.618	.476	-0.060
31	12.00	.756	0.694	.524	0.060
32	12.22	.780	0.774	.571	0.180
33	12.35	.805	0.859	.619	0.303
34	12.59	.829	0.951	.667	0.431
35	12.85	.854	1.052	.714	0.566
36	13.24	.878	1.165	.762	0.712
37	13.31	.902	1.296	.810	0.876
38	13.70	.927	1.453	.857	1.068
39	14.20	.951	1.657	.905	1.309
40	15.00	.976	1.971	.952	1.668

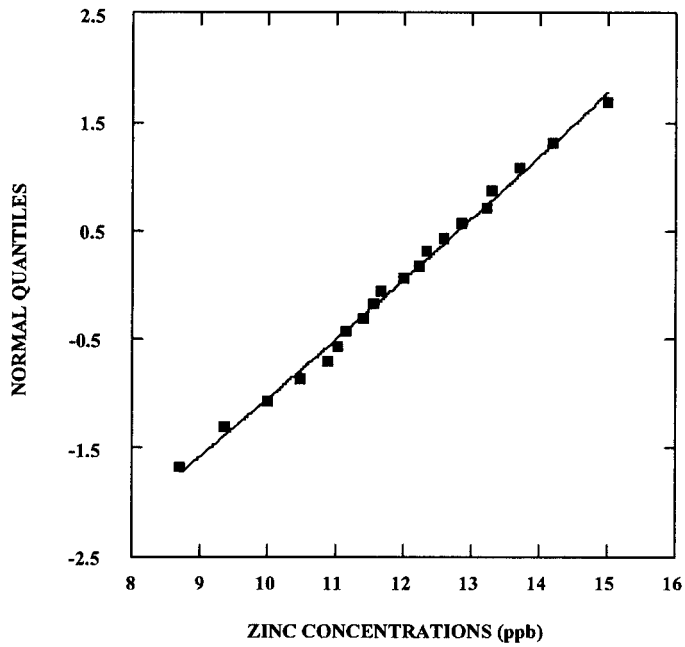
Step 4. Plot the detected zinc concentrations versus each set of probabilities or Normal quantiles, as per the procedure for constructing Probability Plots (see figures below). The nondetect values should not be plotted. As can be seen from the graphs, the Censored Probability Plot indicates a definite curvature in the tails, especially the lower tail. The Detects-Only Probability Plot, however, is reasonably linear. This visual impression is bolstered by calculation of a Probability Plot Correlation Coefficient for each set of detected values: the Censored Probability Plot has a correlation of r=.969, while the Detects-Only Probability Plot has a correlation of r=.998.

Step 5. Because the Detects-Only Probability Plot is substantially more linear than the Censored Probability Plot, it may be appropriate to consider detects and nondetects as arising from statistically distinct distributions, with nondetects representing "zero" concentrations. Therefore, Aitchison's adjustment may lead to better estimates of the true mean and standard deviation than Cohen's adjustment for censored data.

CENSORED PROBABILITY PLOT



DETECTS-ONLY PROBABILITY PLOT



2.2.2 Aitchison's Adjustment

To actually compute Aitchison's adjustment (Aitchison, 1955), it is assumed that the detected samples follow an underlying Normal distribution. If the detects are Lognormal, compute Aitchison's adjustment on the logarithms of the data instead. Let d=# nondetects and let n=total # of samples (detects and nondetects combined). Then if $\bar{x}^*$ and s^* denote respectively the sample mean and standard deviation of the detected values, the adjusted overall mean can be estimated as

$$\hat{\mu} = \left(1 - \frac{d}{n}\right)\bar{x}^*$$

and the adjusted overall standard deviation may be estimated as the square root of the quantity

$$\hat{\sigma}^2 = \frac{n-(d+1)}{n-1}(s^*)^2 + \frac{d}{n}\left(\frac{n-d}{n-1}\right)(\bar{x}^*)^2$$

The general formula for a parametric statistical interval adjusted for nondetects by Aitchison's method is given by $\hat{\mu} \pm K \cdot \hat{\sigma}$, with κ depending on the type of interval being constructed.

EXAMPLE 9

In Example 8, it was determined that Aitchison's adjustment might lead to more appropriate estimates of the true mean and standard deviation than Cohen's adjustment. Use the data in Example 8 to compute Aitchison's adjustment.

SOLUTION

Step 1. The zinc data consists of 20 nondetects and 20 detected values; therefore d=20 and n=40 in the above formulas.

Step 2. Compute the average $\bar{x}^* = 11.891$ and the standard deviation $s^* = 1.595$ of the set of detected values.

Step 3. Use the formulas for Aitchison's adjustment to compute estimates of the true mean and standard deviation:

$$\hat{\mu} = \left(1 - \frac{20}{40}\right) \times 11.891 = 5.95$$

$$\boxed{\hat{\sigma}^2 = \left(\frac{40-21}{39}\right)(1.595)^2 + \left(\frac{20}{40}\right)\left(\frac{20}{39}\right)(11.891)^2 = 37.495 \Rightarrow \hat{\sigma} = 6.12}$$

If Cohen's adjustment is mistakenly computed on these data instead, with a detection limit of 7 ppb,the estimates become $\hat{\ } = 7.63$ and $\hat{\ } = 4.83$. Thus, the choice of adjustment can have a significant impact on the upper limits computed for statistical intervals.

2.2.3 More Than 50% Nondetects

If more than 50% but less than 90% of the samples are nondetect or the assumptions of Cohen's and Aitchison's methods cannot be justified, parametric statistical intervals should be abandoned in favor of non-parametric alternatives (see **Section 3** below). Nonparametric statistical intervals are easy to construct and apply to ground water data measurements, and no special steps need be taken to handle nondetects.

When 90% or more of the data values are nondetect (as often occurs when measuring volatile organic compounds [VOCs] in ground water, for instance), the detected samples can often be modeled as "rare events" by using the Poisson distribution. The Poisson model describes the behavior of a series of independent events over a large number of trials, where the probability of occurrence is low but stays constant from trial to trial. The Poisson model is similar to the Binomial model in that both models represent "counting processes." In the Binomial case,

nondetects are counted as 'misses' or zeroes and detects are counted (regardless of contamination level) as 'hits' or ones; in the case of the Poisson, each particle or molecule of contamination is counted separately but cumulatively, so that the counts for detected samples with high concentrations are larger than counts for samples with smaller concentrations. As Gibbons (1987, p. 574) has noted, it can be postulated

> ...that the number of molecules of a particular compound out of a much larger number of molecules of water is the result of a Poisson process. For example, we might consider 12 ppb of benzene to represent a count of 12 units of benzene for every billion units examined. In this context, Poisson's approach is justified in that the number of units (i.e., molecules) is large, and the probability of the occurrence (i.e., a molecule being classified as benzene) is small.

For a detect with concentration of 50 ppb, the Poisson count would be 50. Counts for nondetects can be taken as zero or perhaps equal to half the detection limit (e.g., if the detection limit were 10 ppb, the Poisson count for that sample would be 5). Unlike the Binomial (Test of Proportions) model, the Poisson model has the ability to utilize the <u>magnitudes</u> of detected concentrations in statistical tests.

The Poisson distribution is governed by the average rate of occurrence, λ, which can be estimated by summing the Poisson counts of all samples in the background pool of data and dividing by the number of samples in the pool. Once the average rate of occurrence has been estimated, the formula for the Poisson distribution is given by

$$\Pr\{X = x\} = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

where x represents the Poisson count and λ represents the average rate of occurrence. To use the Poisson distribution to predict concentration values at downgradient wells, formulas for constructing Poisson Prediction and Tolerance limits are given below.

2.2.4 Poisson Prediction Limits

To estimate a Prediction limit at a particular well using the Poisson model, the approach described by Gibbons (1987b) and based on the work of Cox and Hinkley (1974) can be used. In this case, an upper limit is estimated for an interval that will contain <u>all</u> of k future measurements of an analyte with confidence level $1-\alpha$, given n previous background measurements.

To do this, let T_n represent the sum of the Poisson counts of n background samples. The goal is to predict T_k^*, representing the total Poisson count of the next k sample measurements. As Cox and Hinkley show, if T_n has a Poisson distribution with mean μ and if no contamination has occurred, it is reasonable to assume that T_k^* will also have a Poisson distribution but with mean $c\mu$, where c depends on the number of future measurements being predicted.

In particular, Cox and Hinckley demonstrate that the quantity

$$\left[T_k^* - \frac{c(T_n + T_k^*)}{(1+c)}\right]^2 \Big/ \frac{c(T_n + T_k^*)}{(1+c)^2}$$

has an approximate standard Normal distribution. From this relation, an upper prediction limit for T_k^* is calculated by Gibbons to be approximately

$$T_k^* = cT_n + \frac{ct^2}{2} + ct\sqrt{T_n\left(1+\frac{1}{c}\right)+\frac{t^2}{4}}$$

where $t=t_{n-1,\alpha}$ is the upper $(1-\alpha)$ percentile of the Student's t distribution with (n-1) degrees of freedom. The quantity c in the above formulas may be computed as k/n, where, as noted, k is the number of future samples being predicted.

EXAMPLE 10

Use the following benzene data from six background wells to estimate an upper 99% Poisson Prediction limit for the next four measurements from a single downgradient well.

	Benzene Concentrations (ppb)					
Month	Well 1	Well 2	Well 3	Well 4	Well 5	Well 6
1	<2	<2	<2	<2	<2	<2
2	<2	<2	<2	15.0	<2	<2
3	<2	<2	<2	<2	<2	<2
4	<2	12.0	<2	<2	<2	<2
5	<2	<2	<2	<2	<2	10.0
6	<2	<2	<2	<2	<2	<2

SOLUTION

Step 1. Pooling the background data yields n=36 samples, of which, 33 (92%) are nondetect. Because the rate of detection is so infrequent (i.e., <10%), a Poisson-based Prediction limit may be appropriate. Since four future measurements are to be predicted, k=4, and hence, c=k/n=1/9.

Step 2. Set each nondetect to half the detection limit or 1 ppb. Then compute the Poisson count of the sum of all the background samples, in this case, T_n=33(1)+(12.0+15.0+10.0) = 70.0. To calculate an upper 99% Prediction limit, the upper 99th percentile of the t-distribution with (n-1)=35 degrees of freedom must be taken from a reference table, namely $t_{35,.01}$=2.4377.

Step 3. Using Gibbons' formula above, calculate the upper Prediction limit as:

$$T_k^* = \frac{1}{9}(70) + \frac{(2.4377)^2}{2(9)} + \frac{2.4377}{9}\sqrt{70(1+9) + \frac{(2.4377)^2}{4}} = 15.3 \text{ ppb}$$

Step 4. To test the upper Prediction limit, the Poisson count of the <u>sum</u> of the next four downgradient wells should be calculated. If this sum is greater than 15.3 ppb, there is significant evidence of contamination at the downgradient well. If not, the well may be regarded as clean until the next testing period.

The procedure for generating Poisson prediction limits is somewhat flexible. The value k above, for instance, need not represent multiple samples from a single well. It could also denote a collection of single samples from k distinct wells, all of which are assumed to follow the same Poisson distribution in the absence of contamination. The Poisson distribution also has the desirable property that the sum of several Poisson variables also has a Poisson distribution, even if the individual components are not identically distributed. Because of this, Gibbons (1987b) has suggested that if several analytes (e.g., different VOCs) can all be modeled via the Poisson distribution, the combined sum of the Poisson counts of all the analytes will also have a Poisson distribution, meaning that a single prediction limit could be estimated for the combined group of analytes, thus reducing the necessary number of statistical tests.

A major drawback to Gibbons' proposal of establishing a combined prediction limit for several analytes is that if the limit is exceeded, it will not be clear which analyte is responsible for "triggering" the test. In part this problem explains why the ground-water monitoring regulations mandate that each analyte be tested separately. Still, if a large number of analytes must be regularly tested and the detection rate is quite low, the overall facility-wide false positive rate may be unacceptably high. To remedy this situation, it is probably wisest to do enough initial testing

of background and facility leachate and waste samples to determine those specific parameters present at levels substantially greater than background. By limiting monitoring and statistical tests to a few parameters meeting the above conditions, it should be possible to contain the overall facility-wide false positive rate while satisfying the regulatory requirements and assuring reliable identification of ground-water contamination if it occurs.

Though quantitative information on a suite of VOCs may be automatically generated as a consequence of the analytical method configuration (e.g., SW-846 method 8260 can provide quantitative results for approximately 60 different compounds), it is usually unnecessary to designate all of these compounds as leak detection indicators. Such practice generally aggravates the problem of many comparisons and results in elevated false positive rates for the facility as a whole. This makes accurate statistical testing especially difficult. EPA therefore recommends that the results of leachate testing or the waste analysis plan serve as the primary basis for designating reliable leak detection indicator parameters.

2.2.5 Poisson Tolerance Limits

To apply an upper Tolerance limit using the Poisson model to a group of downgradient wells, the approach described by Gibbons (1987b) and based on the work of Zacks (1970) can be taken. In this case, if no contamination has occurred, the estimated interval upper limit will contain a large fraction of all measurements from the downgradient wells, often specified at 95% or more.

The calculations involved in deriving Poisson Tolerance limits can seem non-intuitive, primarily because the argument leading to a mathematically rigorous Tolerance limit is complicated. The basic idea, however, uses the fact that if each individual measurement follows a common Poisson distribution with rate parameter, λ, the sum of n such measurements will also follow a Poisson distribution, this time with rate $n\lambda$.

Because the Poisson distribution has the property that its true mean is equal to the rate parameter λ, the concentration sum of n background samples can be manipulated to estimate this rate. But since we know that the distribution of the concentration sum is also Poisson, the possible values of λ can actually be narrowed to within a small range with fixed confidence probability (γ).

For each "possible" value of λ in this confidence range, one can compute the percentile of the Poisson distribution with rate λ that would lie above, say, 95% of all future downgradient measurements. By setting as the "probable" rate, that λ which is greater than all but a small percentage α of the most extreme possible λ's, given the values of n background samples, one can compute an upper tolerance limit with, say, 95% coverage and $(1-\alpha)$% confidence.

To actually make these computations, Zacks (1970) shows that the most probable rate λ can be calculated approximately as

$$\lambda_{T_n} = \frac{1}{2n}\chi^2_{\gamma}[2T_n + 2]$$

where as before T_n represents the Poisson count of the sum of n background samples (setting nondetects to half the method detection limit), and

$$\chi^2_{\gamma}[2T_n + 2]$$

represents the γ percentile of the Chi-square distribution with $(2T_n+2)$ degrees of freedom.

To find the upper Tolerance limit with β% coverage (e.g., 95%) once a probable rate λ has been estimated, one must compute the Poisson percentile that is larger than β% of all possible measurements from that distribution, that is, the β% quantile of the Poisson distribution with mean rate λ_{Tn}, denoted by $P^{-1}(\beta,\lambda_{Tn})$. Using a well-known mathematical relationship between the Poisson and Chi-square distributions, finding the β% quantile of the Poisson amounts to determining the least positive integer k such that

$$\chi^2_{1-\beta}[2k+2] \geq 2\lambda_{T_n}$$

where, as above, the quantity [2k+2] represents the degrees of freedom of the Chi-square distribution. By calculating two times the estimated probable rate λ_{Tn} on the right-hand-side of the above inequality, and then finding the smallest degrees of freedom so that the (1-β)% percentile of the Chi-square distribution is bigger than $2\lambda_{Tn}$, the upper tolerance limit k can be determined fairly easily.

Once the upper tolerance limit, k, has been estimated, it will represent an upper Poisson Tolerance limit having approximately β% coverage with γ% confidence in all comparisons with downgradient well measurements.

EXAMPLE 11

Use the benzene data of Example 10 to estimate an upper Poisson Tolerance limit with 95% coverage and 95% confidence probability.

SOLUTION

Step 1. The benzene data consist of 33 nondetects with detection limit equal to 2 ppb and 3 detected values for a total of n=36. By setting each nondetect to half the detection limit as before, one finds a total Poisson count of the sum equal to T_n=70.0. It is also known that the desired confidence probability is γ=.95 and the desired coverage is β=.95.

Step 2. Based on the observed Poisson count of the sum of background samples, estimate the probable occurrence rate λ_{Tn} using Zacks' formula above as

$$\lambda_{T_n} = \frac{1}{2n}\chi^2_{\gamma}[2T_n + 2] = \frac{1}{72}\chi^2_{.95}[142] = 2.37$$

Step 3. Compute twice the probable occurrence rate as $2\lambda_{Tn}$=4.74. Now using a Chi-square table, find the smallest degrees of freedom (df), k, such that

$$\chi^2_{.05}[2k + 2] \geq 4.74$$

Since the 5th percentile of the Chi-square distribution with 12 df equals 5.23 (but only 4.57 with 11 df), it is seen that (2k+2)=12, leading to k=5. Therefore, the upper Poisson Tolerance limit is estimated as k=5 ppb.

Step 4. Because the estimated upper Tolerance limit with 95% coverage equals 5 ppb, any detected value among downgradient samples greater than 5 ppb may indicate possible evidence of contamination.

3. NON-PARAMETRIC COMPARISON OF COMPLIANCE WELL DATA TO BACKGROUND

When concentration data from several compliance wells are to be compared with concentration data from background wells, one basic approach is analysis of variance (ANOVA). The ANOVA technique is used to test whether there is statistically significant evidence that the mean concentration of a constituent is higher in one or more of the compliance wells than the baseline provided by background wells. Parametric ANOVA methods make two key assumptions: 1) that the data residuals are Normally distributed and 2) that the group variances are all approximately equal. The steps for calculating a parametric ANOVA are given in the Interim Final Guidance (pp. 5-6 to 5-14).

If either of the two assumptions crucial to a parametric ANOVA is grossly violated, it is recommended that a non-parametric test be conducted using the ranks of the observations rather than the original observations themselves. The Interim Final Guidance describes the Kruskal-Wallis test when three or more well groups (including background data, see pp. 5-14 to 5-20) are being compared. However, the Kruskal-Wallis test is not amenable to two-group comparisons, say of one compliance well to background data. In this case, the Wilcoxon Rank-Sum procedure (also known as the Mann-Whitney U Test) is recommended and explained below. Since most situations will involve the comparison of at least two downgradient wells with background data, the Kruskal-Wallis test is presented first with an additional example.

3.1 KRUSKAL-WALLIS TEST

When the assumptions used in a parametric analysis of variance cannot be verified, e.g., when the original or transformed residuals are not approximately Normal in distribution or have significantly different group variances, an analysis can be performed using the ranks of the observations. Usually, a non-parametric procedure will be needed when a substantial fraction of the measurements are below detection (more than 15 percent), since then the above assumptions are difficult to verify.

The assumption of independence of the residuals is still required. Under the null hypothesis that there is no difference among the groups, the observations are assumed to come from identical distributions. However, the form of the distribution need not be specified.

A non-parametric ANOVA can be used in any situation that the parametric analysis of variance can be used. However, because the ranks of the data are being used, the minimum sample sizes for the groups must be a little larger. A useful rule of thumb is to require a minimum of three well groups with at least four observations per group before using the Kruskal-Wallis procedure.

Non-parametric procedures typically need a few more observations than parametric procedures for two reasons. On the one hand, non-parametric tests make fewer assumptions concerning the distribution of the data and so more data is often needed to make the same judgment that would be rendered by a parametric test. Also, procedures based on ranks have a discrete distribution (unlike the continuous distributions of parametric tests). Consequently, a larger sample size is usually needed to produce test statistics that will be significant at a specified alpha level such as 5 percent.

The relative efficiency of two procedures is defined as the ratio of the sample sizes needed by each to achieve a certain level of power against a specified alternative hypothesis. As sample sizes get larger, the efficiency of the Kruskal-Wallis test relative to the parametric analysis of variance test approaches a limit that depends on the underlying distribution of the data, but is always at least 86 percent. This means roughly that in the worst case, if 86 measurements are available for a parametric ANOVA, only 100 sample values are needed to have an equivalently powerful Kruskal-Wallis test. In many cases, the increase in sample size necessary to match the power of a parametric ANOVA is much smaller or not needed at all. The efficiency of the Kruskal-Wallis test is 95 percent if the data are really Normal, and can be much larger than 100 percent in other cases (e.g., it is 150 percent if the residuals follow a distribution called the double exponential).

These results concerning efficiency imply that the Kruskal-Wallis test is reasonably powerful for detecting concentration differences despite the fact that the original data have been replaced by their ranks, and can be used even when the data are Normally distributed. When the data are not Normal or cannot be transformed to Normality, the Kruskal-Wallis procedure tends to be more powerful for detecting differences than the usual parametric approach.

3.1.1 Adjusting for Tied Observations

Frequently, the Kruskal-Wallis procedure will be used when the data contain a significant fraction of nondetects (e.g., more than 15 percent of the samples). In these cases, the parametric

assumptions necessary for the usual one-way ANOVA are difficult or impossible to verify, making the non-parametric alternative attractive. However, the presence of nondetects prevents a unique ranking of the concentration values, since nondetects are, up to the limit of measurement, all tied at the same value.

To get around this problem, two steps are necessary. First, in the presence of ties (e.g., nondetects), all tied observations should receive the same rank. This rank (sometimes called the midrank (Lehmann, 1975)) is computed as the average of the ranks that would be given to a group of ties if the tied values actually differed by a tiny amount and could be ranked uniquely. For example, if the first four ordered observations are all nondetects, the midrank given to each of these samples would be equal to (1+2+3+4)/4=2.5. If the next highest measurement is a unique detect, its rank would be 5 and so on until all observations are appropriately ranked.

The second step is to compute the Kruskal-Wallis statistic as described in the Interim Final Guidance, using the midranks computed for the tied values. Then an adjustment to the Kruskal-Wallis statistic must be made to account for the presence of ties. This adjustment is described on page 5-17 of the Interim Final Guidance and requires computation of the formula:

$$H' = \frac{H}{1 - \left(\Sigma_{i=1}^{g} \frac{t_i^3 - t_i}{N^3 - N} \right)}$$

where g equals the number of groups of distinct tied observations and t_i is the number of observations in the ith tied group.

EXAMPLE 12

Use the non-parametric analysis of variance on the following data to determine whether there is evidence of contamination at the monitoring site.

	Toluene Concentration (ppb)				
	Background Wells		Compliance Wells		
Month	Well 1	Well 2	Well 3	Well 4	Well 5
1	<5	<5	<5	<5	<5
2	7.5	<5	12.5	13.7	20.1
3	<5	<5	8.0	15.3	35.0
4	<5	<5	<5	20.2	28.2

5	6.4	<5	11.2	25.1	19.0

SOLUTION

Step 1. Compute the overall percentage of nondetects. In this case, nondetects account for 48 percent of the data. The usual parametric analysis of variance would be inappropriate. Use the Kruskal-Wallis test instead, pooling both background wells into one group and treating each compliance well as a separate group.

Step 2. Compute ranks for all the data including tied observations (e.g., nondetects) as in the following table. Note that each nondetect is given the same midrank, equal to the average of the first 12 unique ranks.

	Toluene Ranks				
	Background Wells		Compliance Wells		
Month	Well 1	Well 2	Well 3	Well 4	Well 5
1	6.5	6.5	6.5	6.5	6.5
2	14	6.5	17	18	21
3	6.5	6.5	15	19	25
4	6.5	6.5	6.5	22	24
5	13	6.5	16	23	20
Rank Sum	R_b=79		R_3=61	R_4=88.5	R_5=96.5
Rank Mean	$\bar{R}_b$=7.9		$\bar{R}_3$=12.2	$\bar{R}_4$=17.7	$\bar{R}_5$=19.3

Step 3. Calculate the sums of the ranks in each group (R_i) and the mean ranks in each group ($\bar{R}_i$). These results are given above.

Step 4. Compute the Kruskal-Wallis statistic H using the formula on p. 5-15 of the Interim Final Guidance

$$H = \left[\frac{12}{N(N+1)} \sum_{i=1}^{K} \frac{R_i^2}{N_i} \right] - 3(N+1)$$

where N=total number of samples, N_i=number of samples in ith group, and K=number of groups. In this case, N=25, K=4, and H can be computed as

$$H = \frac{12}{25*26}\left[\frac{79^2}{10} + \frac{61^2}{5} + \frac{88.5^2}{5} + \frac{96.5^2}{5}\right] - 78 = 10.56.$$

Step 5. Compute the adjustment for ties. There is only one group of distinct tied observations, containing 12 samples. Thus, the adjusted Kruskal-Wallis statistic is given by:

$$H' = \frac{10.56}{1 - \left(\frac{12^3 - 12}{25^3 - 25}\right)} = 11.87.$$

Step 6. Compare the calculated value of H′ to the tabulated Chi-square value with (K-1)= (# groups-1)=3 df, $\chi^2_{3,.05}$=7.81. Since the observed value of 11.87 is greater than the Chi-square critical value, there is evidence of significant differences between the well groups. Post-hoc pairwise comparisons are necessary.

Step 7. Calculate the critical difference for compliance well comparisons to the background using the formula on p. 5-16 of the Interim Final Guidance document. Since the number of samples at each compliance well is four, the same critical difference can be used for each comparison, namely,

$$C_i = z_{.05/3}\sqrt{\frac{25 \cdot 26}{12}\left(\frac{1}{10} + \frac{1}{5}\right)} = 8.58$$

Step 8. Form the differences between the average ranks of each compliance well and the background and compare these differences to the critical value of 8.58.

Well 3: $\bar{R}_3 - \bar{R}_b = 12.2\text{-}7.9 = 4.3$

Well 4: $\bar{R}_4 - \bar{R}_b = 17.7\text{-}7.9 = 9.8$

Well 5: $\bar{R}_5 - \bar{R}_b = 19.3\text{-}7.9 = 11.4$

Since the average rank differences at wells 4 and 5 exceed the critical difference, there is significant evidence of contamination at wells 4 and 5, but not at well 3.

3.2 WILCOXON RANK-SUM TEST FOR TWO GROUPS

When a single compliance well group is being compared to background data and a non-parametric test is needed, the Kruskal-Wallis procedure should be replaced by the Wilcoxon Rank-Sum test (Lehmann, 1975; also known as the two-sample Mann-Whitney U test). For most ground-water applications, the Wilcoxon test should be used whenever the proportion of nondetects in the combined data set exceeds 15 percent. However, to provide valid results, do

not use the Wilcoxon test unless the compliance well and background data groups both contain at least four samples each.

To run the Wilcoxon Rank-Sum Test, use the following algorithm. Combine the compliance and background data and rank the ordered values from 1 to N. Assume there are n compliance samples and m background samples so that N=m+n. Denote the ranks of the compliance samples by C_i and the ranks of the background samples by B_i. Then add up the ranks of the compliance samples and subtract n(n+1)/2 to get the Wilcoxon statistic W:

$$W = \Sigma_{i=1}^{n} C_i - \frac{1}{2} n(n+1).$$

The rationale of the Wilcoxon test is that if the ranks of the compliance data are quite large relative to the background ranks, then the hypothesis that the compliance and background values came from the same population should be rejected. Large values of the statistic W give evidence of contamination at the compliance well site.

To find the critical value of W, a Normal approximation to its distribution is used. The expected value and standard deviation of W under the null hypothesis of no contamination are given by the formulas

$$E(W) = \frac{1}{2} mn; \qquad SD(W) = \sqrt{\frac{1}{12} mn(N+1)}$$

An approximate Z-score for the Wilcoxon Rank-Sum Test then follows as:

$$Z \approx \frac{W - E(W) - \frac{1}{2}}{SD(W)}.$$

The factor of 1/2 in the numerator serves as a continuity correction since the discrete distribution of the statistic W is being approximated by the continuous Normal distribution.

Once an approximate Z-score has been computed, it may be compared to the upper 0.01 percentile of the standard Normal distribution, $z_{.01}$=2.326, in order to determine the statistical significance of the test. If the observed Z-score is greater than 2.326, the null hypothesis may be

rejected at the 1 percent significance level, suggesting that there is significant evidence of contamination at the compliance well site.

EXAMPLE 13

The table below contains copper concentration data (ppb) found in water samples at a monitoring facility. Wells 1 and 2 are background wells and well 3 is a single compliance well suspected of contamination. Calculate the Wilcoxon Rank-Sum Test on these data.

	Copper Concentration (ppb)		
	Background		Compliance
Month	Well 1	Well 2	Well 3
1	4.2	5.2	9.4
2	5.8	6.4	10.9
3	11.3	11.2	14.5
4	7.0	11.5	16.1
5	7.3	10.1	21.5
6	8.2	9.7	17.6

SOLUTION

Step 1. Rank the N=18 observations from 1 to 18 (smallest to largest) as in the following table.

	Ranks of Copper Concentrations		
	Background		Compliance
Month	Well 1	Well 2	Well 3
1	1	2	8
2	3	4	11
3	13	12	15
4	5	14	16
5	6	10	18
6	7	9	17

Step 2. Compute the Wilcoxon statistic by adding up the compliance well ranks and subtracting $n(n+1)/2$, so that $W=85-21=64$.

Step 3. Compute the expected value and standard deviation of W.

$$E(W) = \frac{1}{2}mn = 36$$

$$SD(W) = \sqrt{\frac{1}{12}mn(N+1)} = \sqrt{114} = 10.677$$

Step 4. Form the approximate Z-score.

$$Z \approx \frac{W - E(W) - \frac{1}{2}}{SD(W)} = \frac{64 - 36 - 0.5}{10.677} = 2.576$$

Step 5. Compare the observed Z-score to the upper 0.01 percentile of the Normal distribution. Since $Z=2.576>2.326=z_{.01}$, there is significant evidence of contamination at the compliance well at the 1 percent significance level.

3.2.1 Handling Ties in the Wilcoxon Test

Tied observations in the Wilcoxon test are handled in similar fashion to the Kruskal-Wallis procedure. First, midranks are computed for all tied values. Then the Wilcoxon statistic is computed as before but with a slight difference. To form the approximate Z-score, an adjustment is made to the formula for the standard deviation of W in order to account for the groups of tied values. The necessary formula (Lehmann, 1975) is:

$$SD^{*}(W) = \sqrt{\frac{mn(N+1)}{12}\left(1 - \Sigma_{i=1}^{g}\frac{t_i^3 - t_i}{N^3 - N}\right)}$$

where, as in the Kruskal-Wallis method, g equals the number of groups of distinct tied observations and t_i represents the number of tied values in the ith group.

4. STATISTICAL INTERVALS: CONFIDENCE, TOLERANCE, AND PREDICTION

Three types of statistical intervals are often constructed from data: Confidence intervals, Tolerance intervals, and Prediction intervals. Though often confused, the interpretations and uses of these intervals are quite distinct. The most common interval encountered in a course on statistics is a Confidence interval for some parameter of the distribution (e.g., the population mean). The interval is constructed from sample data and is thus a random quantity. This means that each set of sample data will generate a different Confidence interval, even though the algorithm for constructing the interval stays the same every time.

A Confidence interval is designed to contain the specified population parameter (usually the mean concentration of a well in ground-water monitoring) with a designated level of confidence or probability, denoted as $1-\alpha$. The interval will fail to include the true parameter in approximately α percent of the cases where such intervals are constructed.

The usual Confidence interval for the mean gives information about the average concentration level at a particular well or group of wells. It offers little information about the highest or most extreme sample concentrations one is likely to observe over time. Often, it is those extreme values one wants to monitor to be protective of human health and the environment. As such, a Confidence interval generally should be used only in two situations for ground-water data analysis: (1) when directly specified by the permit or (2) in compliance monitoring, when downgradient samples are being compared to a Ground-Water Protection Standard (GWPS) representing the average of onsite background data, as is sometimes the case with an Alternate Contaminant Level (ACL) . In other situations it is usually desirable to employ a Tolerance or Prediction interval.

A Tolerance interval is designed to contain a designated proportion of the population (e.g., 95 percent of all possible sample measurements). Since the interval is constructed from sample data, it also is a random interval. And because of sampling fluctuations, a Tolerance interval can contain the specified proportion of the population only with a certain confidence level. Two coefficients are associated with any Tolerance interval. One is the proportion of the population that the interval is supposed to contain, called the coverage. The second is the degree of confidence with which the interval reaches the specified coverage. This is known as the tolerance coefficient. A Tolerance interval with coverage of 95 percent and a tolerance coefficient of 95

percent is constructed to contain, on average, 95 percent of the distribution with a probability of 95 percent.

Tolerance intervals are very useful for ground-water data analysis, because in many situations one wants to ensure that at most a small fraction of the compliance well sample measurements exceed a specific concentration level (chosen to be protective of human health and the environment). Since a Tolerance interval is designed to cover all but a small percentage of the population measurements, observations should very rarely exceed the upper Tolerance limit when testing small sample sizes. The upper Tolerance limit allows one to gauge whether or not too many extreme concentration measurements are being sampled from compliance point wells.

Tolerance intervals can be used in detection monitoring when comparing compliance data to background values. They also should be used in compliance monitoring when comparing compliance data to certain Ground-Water Protection Standards. Specifically, the tolerance interval approach is recommended for comparison with a Maximum Contaminant Level (MCL) or with an ACL if the ACL is derived from health-based risk data.

Prediction intervals are constructed to contain the next sample value(s) from a population or distribution with a specified probability. That is, after sampling a background well for some time and measuring the concentration of an analyte, the data can be used to construct an interval that will contain the next analyte sample or samples (assuming the distribution has not changed). A Prediction interval will thus contain a future value or values with specified probability. Prediction intervals can also be constructed to contain the average of several future observations.

Prediction intervals are probably most useful for two kinds of detection monitoring. The first kind is when compliance point well data are being compared to background values. In this case the Prediction interval is constructed from the background data and the compliance well data are compared to the upper Prediction limits. The second kind is when intrawell comparisons are being made on an uncontaminated well. In this case, the Prediction interval is constructed on past data sampled from the well, and used to predict the behavior of future samples from the same well.

In summary, a Confidence interval usually contains an average value, a Tolerance interval contains a proportion of the population, and a Prediction interval contains one or more future observations. Each has a probability statement or "confidence coefficient" associated with it. For further explanation of the differences between these interval types, see Hahn (1970).

One should note that all of these intervals assume that the sample data used to construct the intervals are Normally distributed. In light of the fact that much ground-water concentration data is better modeled by a Lognormal distribution, it is recommended that tests for Normality be run on the logarithms of the original data before constructing the random intervals. If the data follow the Lognormal model, then the intervals should be constructed using the logarithms of the sample values. In this case, the limits of these intervals should not be compared to the original compliance data or GWPS. Rather, the comparison should involve the logged compliance data or logged GWPS. When neither the Normal or Lognormal models can be justified, a non-parametric version of each interval may be utilized.

4.1 TOLERANCE INTERVALS

In detection monitoring, the compliance point samples are assumed to come from the same distribution as the background values until significant evidence of contamination can be shown. To test this hypothesis, a 95 percent coverage Tolerance interval can be constructed on the background data. The background data should first be tested to check the distributional assumptions. Once the interval is constructed, each compliance sample is compared to the upper Tolerance limit. If any compliance point sample exceeds the limit, the well from which it was drawn is judged to have significant evidence of contamination (note that when testing a large number of samples, the nature of a Tolerance interval practically ensures that a few measurements will be above the upper Tolerance limit, even when no contamination has occurred. In these cases, the offending wells should probably be resampled in order to verify whether or not there is definite evidence of contamination.)

If the Tolerance limit has been constructed using the logged background data, the compliance point samples should first be logged before comparing with the upper Tolerance limit. The steps for computing the actual Tolerance interval in detection monitoring are detailed in the Interim Final Guidance on pp. 5-20 to 5-24. One point about the table of factors κ used to adjust the width of the Tolerance interval is that these factors are designed to provide at least 95% coverage of the population. Applied over many data sets, the average coverage of these intervals will often be close to 98% or more (see Guttman, 1970). To construct a one-sided upper Tolerance interval with average coverage of $(1-\beta)\%$, the κ multiplier can be computed directly with the aid of a Student's t-distribution table. In this case, the formula becomes

$$= t_{n-1,1-\beta}\sqrt{1+\frac{1}{n}}$$

where the t-value represents the (1-β)th upper percentile of the t-distribution with (n-1) degrees of freedom.

In compliance monitoring, the Tolerance interval is calculated on the compliance point data, so that the upper one-sided Tolerance limit may be compared to the appropriate Ground-Water Protection Standard (i.e., MCL or ACL). If the upper Tolerance limit exceeds the fixed standard, and especially if the Tolerance limit has been constructed to have an average coverage of 95% as described above, there is significant evidence that as much as 5 percent or more of all the compliance well measurements will exceed the limit and consequently that the compliance point wells are in violation of the facility permit. The algorithm for computing Tolerance limits in compliance monitoring is given on pp. 6-11 to 6-15 of the Interim Final Guidance.

EXAMPLE 14

The table below contains data that represent chrysene concentration levels (ppb) found in water samples obtained from the five compliance wells at a monitoring facility. Compute the upper Tolerance limit at each well for an average of 95% coverage with 95% confidence and determine whether there is evidence of contamination. The alternate concentration limit (ACL) is 80 ppb.

	Chrysene Concentration (ppb)				
Month	Well 1	Well 2	Well 3	Well 4	Well 5
1	19.7	10.2	68.0	26.8	47.0
2	39.2	7.2	48.9	17.7	30.5
3	7.8	16.1	30.1	31.9	15.0
4	12.8	5.7	38.1	22.2	23.4
Mean	19.88	9.80	46.28	24.65	28.98
SD	13.78	4.60	16.40	6.10	13.58

SOLUTION

Step 1. Before constructing the tolerance intervals, check the distributional assumptions. The algorithm for a parametric Tolerance interval assumes that the data used to compute the interval are Normally distributed. Because these data are more likely to be Lognormal in distribution than Normal, check the assumptions on the logarithms of the original data given in the table below. Since each well has only four observations, Probability Plots are not likely to be informative. The Shapiro-Wilk or Probability Plot Correlation

Coefficient tests can be run, but in this example only the Skewness Coefficient is examined to ensure that gross departures from Lognormality are not missed.

	Logged Chrysene Concentration [log(ppb)]				
Month	Well 1	Well 2	Well 3	Well 4	Well 5
1	2.98	2.32	4.22	3.29	3.85
2	3.67	1.97	3.89	2.87	3.42
3	2.05	2.78	3.40	3.46	2.71
4	2.55	1.74	3.64	3.10	3.15
Mean	2.81	2.20	3.79	3.18	3.28
SD	0.68	0.45	0.35	0.25	0.48

Step 2. The Skewness Coefficients for each well are given in the following table. Since none of the coefficients is greater than 1 in absolute value, approximate Lognormality (that is, Normality of the logged data) is assumed for the purpose of constructing the tolerance intervals.

Well	Skewness	|Skewness|
1	.210	.210
2	.334	.334
3	.192	.192
4	-.145	.145
5	-.020	.020

Step 3. Compute the tolerance interval for each compliance well using the logged concentration data. The means and SDs are given in the second table above.

Step 4. The tolerance factor for a one-sided Normal tolerance interval with an <u>average</u> of 95% coverage with 95% probability and n=4 observations is given by

$$= t_{3,.05}\sqrt{1+\frac{1}{4}} = 2.631$$

The upper tolerance limit is calculated below for each of the five wells.

Well 1 2.81+2.631(0.68)= 4.61 log(ppb)

Well 2	2.20+2.631(0.45)= 3.38 log(ppb)
Well 3	3.79+2.631(0.35)= 4.71 log(ppb)
Well 4	3.18+2.631(0.25)= 3.85 log(ppb)
Well 5	3.28+2.631(0.48)= 4.54 log(ppb)

Step 5. Compare the upper tolerance limit for each well to the logarithm of the ACL, that is log(80)=4.38. Since the upper tolerance limits for wells 1, 3, and 5 exceed the logged ACL of 4.38 log(ppb), there is evidence of chrysene contamination in wells 1, 3, and 5.

4.1.1 Non-parametric Tolerance Intervals

When the assumptions of Normality and Lognormality cannot be justified, especially when a significant portion of the samples are nondetect, the use of non-parametric tolerance intervals should be considered. The upper Tolerance limit in a non-parametric setting is usually chosen as an order statistic of the sample data (see Guttman, 1970), commonly the maximum value or maybe the second largest value observed. As a consequence, non-parametric intervals should be constructed only from wells that are not contaminated. Because the maximum sample value is often taken as the upper Tolerance limit, non-parametric Tolerance intervals are very easy to construct and use. The sample data must be ordered, but no ranks need be assigned to the concentration values other than to determine the largest measurements. This also means that nondetects do not have to be uniquely ordered or handled in any special manner.

One advantage to using the maximum concentration instead of assigning ranks to the data is that non-parametric intervals (including Tolerance intervals) are sensitive to the actual magnitudes of the concentration data. Another plus is that unless all the sample data are nondetect, the maximum value will be a detected concentration, leading to a well-defined upper Tolerance limit.

Once an order statistic of the sample data (e.g., the maximum value) is chosen to represent the upper tolerance limit, Guttman (1970) has shown that the coverage of the interval, constructed repeatedly over many data sets, has a Beta probability density with cumulative distribution

$$I_t(n-m+1,m) = \int_0^t \frac{\Gamma(n+1)}{\Gamma(n-m+1)\Gamma(m)} u^{n-m}(1-u)^{m-1}\,du$$

where n=# samples in the data set and m=[(n+1)−(rank of upper tolerance limit value)]. If the maximum sample value is selected as the tolerance limit, its rank is equal to n and so m=1. If the second largest value is chosen as the limit, its rank would be equal to (n−1) and so m=2.

Since the Beta distribution is closely related to the more familiar Binomial distribution, Guttman has shown that in order to construct a non-parametric tolerance interval with at least β% coverage and (1-α) confidence probability, the number of (background) samples must be chosen such that

$$\sum_{t=m}^{n}\binom{n}{t}(1-\beta)^{t}\beta^{n-t} \geq 1-\alpha$$

Table A-6 in Appendix A provides the minimum coverage levels with 95% confidence for various choices of n, using either the maximum sample value or the second largest measurement as the tolerance limit. As an example, with 16 background measurements, the minimum coverage is β=83% if the maximum background value is designated as the upper Tolerance limit and β=74% if the Tolerance limit is taken to be the second largest background value. In general, Table A-6 illustrates that <u>if the underlying distribution of concentration values is unknown, more background samples are needed compared to the parametric setting in order to construct a tolerance interval with sufficiently high coverage</u>. Parametric tolerance intervals do not require as many background samples precisely because the form of the underlying distribution is assumed to be known.

Because the coverage of the above non-parametric Tolerance intervals follows a Beta distribution, it can also be shown that the <u>average</u> (not the <u>minimum</u> as discussed above) level of coverage is equal to 1−[m/(n+1)] (see Guttman, 1970). In particular, when the maximum sample value is chosen as the upper tolerance limit, m=1, and the <u>expected coverage</u> is equal to n/(n+1). This implies that at least 19 background samples are necessary to achieve 95% coverage on average.

EXAMPLE 15

Use the following copper background data to establish a non-parametric upper Tolerance limit and determine if either compliance well shows evidence of copper contamination.

Copper Concentration (ppb)

	Background Wells			Compliance Wells	
Month	Well 1	Well 2	Well 3	Well 4	Well 5
1	<5	9.2	<5		
2	<5	<5	5.4		
3	7.5	<5	6.7		
4	<5	6.1	<5		
5	<5	8.0	<5	6.2	<5
6	<5	5.9	<5	<5	<5
7	6.4	<5	<5	7.8	5.6
8	6.0	<5	<5	10.4	<5

SOLUTION

Step 1. Examine the background data in Wells 1, 2, and 3 to determine that the maximum observed value is 9.2 ppb. Set the 95% confidence upper Tolerance limit equal to this value. Because 24 background samples are available, Table A-6 indicates that the minimum coverage is equal to 88% (the expected average coverage, however, is equal to 24/25=96%). To increase the coverage level, more background samples would have to be collected.

Step 2. Compare each sample in compliance Wells 4 and 5 to the upper Tolerance limit. Since none of the measurements at Well 5 is above 9.2 ppb, while one sample from Well 4 is above the limit, conclude that there is significant evidence of copper contamination at Well 4 but not Well 5.

4.2 PREDICTION INTERVALS

When comparing background data to compliance point samples, a Prediction interval can be constructed on the background values. If the distributions of background and compliance point data are really the same, all the compliance point samples should be contained below the upper Prediction interval limit. Evidence of contamination is indicated if one or more of the compliance samples lies above the upper Prediction limit.

With intrawell comparisons, a Prediction interval can be computed on past data to contain a specified number of future observations from the same well, provided the well has not been previously contaminated. If any one or more of the future samples falls above the upper Prediction limit, there is evidence of recent contamination at the well. The steps to calculate parametric Prediction intervals are given on pp. 5-24 to 5-28 of the Interim Final Guidance.

EXAMPLE 16

The data in the table below are benzene concentrations measured at a groundwater monitoring facility. Calculate the Prediction interval and determine whether there is evidence of contamination.

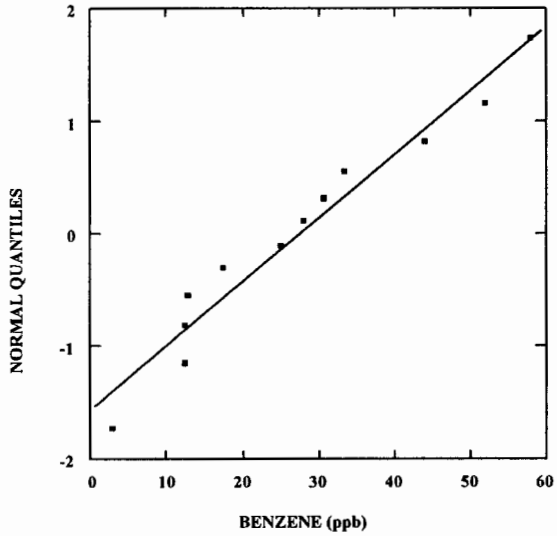
Background Well Data		Compliance Well Data	
Sampling Date	Benzene Concentration (ppb)	Sampling Date	Benzene Concentration (ppb)
Month 1	12.6	Month 4	48.0
	30.8		30.3
	52.0		42.5
	28.1		15.0
Month 2	33.3		
	44.0		n=4
	3.0		Mean=33.95
	12.8		SD=14.64
Month 3	58.1	Month 5	47.6
	12.6		3.8
	17.6		2.6
	25.3		51.9
	n=12		n=4
	Mean=27.52		Mean=26.48
	SD=17.10		SD=26.94

SOLUTION

Step 1. First test the background data for approximate Normality. Only the background data are included since these values are used to construct the Prediction interval.

Step 2. A Probability Plot of the 12 background values is given below. The plot indicates an overall pattern that is reasonably linear with some modest departures from Normality. To further test the assumption of Normality, run the Shapiro-Wilk test on the background data.

PROBABILITY PLOT



Step 3. List the data in ascending and descending order as in the following table. Also calculate the differences $x_{(n-i+1)}-x_{(i)}$ and multiply by the coefficients a_{n-i+1} taken from Table A-1 to get the components of vector b_i used to calculate the Shapiro-Wilk statistic (W).

i	$x_{(i)}$	$x_{(n-i+1)}$	a_{n-i+1}	b_i
1	3.0	58.1	0.548	30.167
2	12.6	52.0	0.333	13.101
3	12.6	44.0	0.235	7.370
4	12.8	33.3	0.159	3.251
5	17.6	30.8	0.092	1.217
6	25.3	28.1	0.030	0.085
7	28.1	25.3		b=55.191
8	30.8	17.6		
9	33.3	12.8		
10	44.0	12.6		
11	52.0	12.6		
12	58.1	3.0		

Step 4. Sum the components b_i in column 5 to get quantity b. Compute the standard deviation of the background benzene values. Then the Shapiro-Wilk statistic is given as

$$W = \left[\frac{b}{SD\sqrt{n-1}}\right]^2 = \left[\frac{55.191}{17.101\sqrt{11}}\right]^2 = 0.947.$$

Step 5. The critical value at the 5% level for the Shapiro-Wilk test on 12 observations is 0.859. Since the calculated value of W=0.947 is well above the critical value, there is no evidence to reject the assumption of Normality.

Step 6. Compute the Prediction interval using the original background data. The mean and standard deviation of the 12 background samples are given by 27.52 ppb and 17.10 ppb, respectively.

Step 7. Since there are two future months of compliance data to be compared to the Prediction limit, the number of future sampling periods is k=2. At each sampling period, a mean of four independent samples will be computed, so m=4 in the prediction interval formula (see Interim Final Guidance, p. 5-25). The Bonferroni t-statistic, $t_{(11,2,.95)}$, with k=2 and 11 df is equivalent to the usual t-statistic at the .975 level with 11 df, i.e., $t_{11,.975}$=2.201.

Step 8. Compute the upper one-sided Prediction limit (UL) using the formula:

$$\overline{X} + t_{(n-1,k,.95)} S\sqrt{\frac{1}{m} + \frac{1}{n}}$$

Then the UL is given by:

$$UL = 27.52 + (17.10)(2.201)\sqrt{\frac{1}{4} + \frac{1}{12}} = 49.25 \text{ ppb.}$$

Step 9. Compare the UL to the compliance data. The means of the four compliance well observations for months 4 and 5 are 33.95 ppb and 26.48 ppb, respectively. Since the mean concentrations for months 4 and 5 are below the upper Prediction limit, there is no evidence of recent contamination at the monitoring facility.

4.2.1 Non-parametric Prediction Intervals

When the parametric assumptions of a Normal-based Prediction limit cannot be justified, often due to the presence of a significant fraction of nondetects, a non-parametric Prediction interval may be considered instead. A non-parametric upper Prediction limit is typically constructed in the same way as a non-parametric upper Tolerance limit, that is, by estimating the limit to be the maximum value of the set of background samples.

The difference between non-parametric Tolerance and Prediction limits is one of interpretation and probability. Given n background measurements and a desired confidence level, a non-parametric Tolerance interval will have a certain coverage percentage. With high probability, the Tolerance interval is designed to miss only a small percentage of the samples from downgradient wells. A Prediction limit, on the other hand, involves the confidence probability that the next future sample or samples will definitely fall below the upper Prediction limit. In this sense, the Prediction limit may be thought of as a 100% coverage Tolerance limit for the next k future samples.

As Guttman (1970) has indicated, the confidence probability associated with predicting that the next single observation from a downgradient well will fall below the upper Prediction limit -- estimated as the maximum background value -- is the <u>same</u> as the <u>expected coverage</u> of a similarly constructed upper Tolerance limit, namely $(1-\alpha)=n/(n+1)$. Furthermore, it can be shown from Gibbons (1991b) that the probability of having k future samples all fall below the upper non-parametric Prediction limit is $(1-\alpha)=n/(n+k)$. Table A-7 in Appendix A lists these confidence levels for various choices of n and k. The false positive rate associated with a single Prediction limit can be computed as one minus the confidence level.

Balancing the ease with which non-parametric upper Prediction limits are constructed is the fact that, given fixed numbers of background samples and future sample values to be predicted, the maximum confidence level associated with the Prediction limit is also fixed. To increase the

level of confidence, the only choices are to 1) decrease the number of future values to be predicted at any testing period, or 2) increase the number of background samples used in the test. Table A-7 can be used along these lines to plan an appropriate sampling strategy so that the false positive rate can be minimized and the confidence probability maximized to a desired level.

EXAMPLE 17

Use the following arsenic data from a monitoring facility to compute a non-parametric upper Prediction limit that will contain the next 2 monthly measurements from a downgradient well and determine the level of confidence associated with the Prediction limit.

	Arsenic Concentrations (ppb)			
	Background Wells			Compliance
Month	Well 1	Well 2	Well 3	Well 4
1	<5	7	<5	
2	<5	6.5	<5	
3	8	<5	10.5	
4	<5	6	<5	
5	9	12	<5	8
6	10	<5	9	14

SOLUTION

Step 1. Determine the maximum value of the background data and use this value to estimate the upper Prediction limit. In this case, the Prediction limit is set to the maximum value of the n=18 samples, or 12 ppb. As is true of non-parametric Tolerance intervals, only uncontaminated wells should be used in the construction of Prediction limits.

Step 2. Compute the confidence level and false positive rate associated with the Prediction limit. Since two future samples are being predicted and n=18, the confidence level is found to be n/(n+k)=18/20=90%. Consequently, the Type I error or false positive rate is equal to (1−.90)=10%. If a lower false positive rate is desired, the number of background samples used in the test must be enlarged.

Step 3. Compare each of the downgradient samples against the upper Prediction limit. Since the value of 14 ppb for month 2 exceeds the limit, conclude that there is significant evidence of contamination at the downgradient well at the 10% level of significance.

4.3 CONFIDENCE INTERVALS

Confidence intervals should only be constructed on data collected during compliance monitoring, in particular when the Ground-Water Protection Standard (GWPS) is an ACL computed from the average of background samples. Confidence limits for the average concentration levels at compliance wells should not be compared to MCLs. Unlike a Tolerance interval, Confidence limits for an average do not indicate how often individual samples will exceed the MCL. Conceivably, the lower Confidence limit for the mean concentration at a compliance well could fall below the MCL, yet 50 percent or more of the individual samples might exceed the MCL. Since an MCL is designed to set an upper bound on the acceptable contamination, this would not be protective of human health or the environment.

When comparing individual compliance wells to an ACL derived from average background levels, a lower one-sided 99 percent Confidence limit should be constructed. If the lower Confidence limit exceeds the ACL, there is significant evidence that the true mean concentration at the compliance well exceeds the GWPS and that the facility permit has been violated. Again, in most cases, a Lognormal model will approximate the data better than a Normal distribution model. It is therefore recommended that the initial data checking and analysis be performed on the logarithms of the data. If a Confidence interval is constructed using logged concentration data, the lower Confidence limit should be compared to the logarithm of the ACL rather than the original GWPS. Steps for computing Confidence intervals are given on pp. 6-3 to 6-11 of the Interim Final Guidance.

5. STRATEGIES FOR MULTIPLE COMPARISONS

5.1 BACKGROUND OF PROBLEM

Multiple comparisons occur whenever more than one statistical test is performed during any given monitoring or evaluation period. These comparisons can arise as a result of the need to test multiple downgradient wells against a pool of upgradient background data or to test several indicator parameters for contamination on a regular basis. Usually the same statistical test is performed in every comparison, each test having a fixed level of confidence ($1-\alpha$), and a corresponding false positive rate, α.

The false positive rate (or Type I error) for an individual comparison is the probability that the test will falsely indicate contamination, i.e., that the test will "trigger," though no contamination has occurred. If ground-water data measurements were always constant in the absence of contamination, false positives would never occur. But ground-water measurements typically vary, either due to natural variation in the levels of background concentrations or to variation in lab measurement and analysis.

Applying the same test to each comparison is acceptable if the number of comparisons is small, but when the number of comparisons is moderate to large the false positive rate associated with the testing network as a whole (that is, across all comparisons involving a separate statistical test) can be quite high. This means that if enough tests are run, there will be a significant chance that at least one test will indicate contamination, even if no actual contamination has occurred. As an example, if the testing network consists of 20 separate comparisons (some combination of multiple wells and/or indicator parameters) and a 99% confidence level Prediction interval limit is used on each comparison, one would expect an overall network-wide false positive rate of over 18%, even though the Type I error for any single comparison is only 1%. This means there is nearly 1 chance in 5 that one or more comparisons will falsely register potential contamination even if none has occurred. With 100 comparisons and the same testing procedure, the overall network-wide false positive rate jumps to more than 63%, adding additional expense to verify the lack of contamination at falsely triggered wells.

To lower the network-wide false positive rate, there are several important considerations. As noted in Section 2.2.4, only those constituents that have been shown to be <u>reliable</u> indicators of potential contamination should be statistically tested on a regular basis. By limiting the number of tested constituents to the most useful indicators, the overall number of statistical comparisons

that must be made can be reduced, lowering the facility-wide false alarm rate. In addition, depending on the hydrogeology of the site, some indicator parameters may need to be tested only at one (or a few adjacent) regulated waste units, as opposed to testing across the entire facility, as long as the permit specifies a common point of compliance, thus further limiting the number of total statistical comparisons necessary.

One could also try to lower the Type I error applied to each individual comparison. Unfortunately, for a given statistical test in general, the lower the false positive rate, the lower the power of the test to detect real contamination at the well. If the statistical power drops too much, real contamination will not be identified when it occurs, creating a situation not protective of the environment or human health. Instead, alternative testing strategies can be considered that specifically account for the number of statistical comparisons being made during any evaluation period. All alternative testing strategies should be evaluated in light of two basic goals:

1. Is the network-wide false positive rate (across all constituents and wells being tested) acceptably low? and

2. Does the testing strategy have adequate statistical power to detect real contamination when it occurs?

To establish a standard recommendation for the network-wide overall false positive rate, it should be noted that for some statistical procedures, EPA specifications mandate that the Type I error for any individual comparison be at least 1%. The rationale for this minimum requirement is motivated by statistical power. For a given test, if the Type I error is set too low, the power of the test will dip below "acceptable" levels. EPA was not able to specify a minimum level of acceptable power within the regulations because to do so would require specification of a minimum difference of environmental concern between the null and alternative hypotheses. Limited current knowledge about the health and/or environmental effects associated with incremental changes in concentration levels of Appendix IX constituents greatly complicates this task. Therefore, minimum false positive rates were adopted for some statistical procedures until more specific guidance could be recommended. EPA's main objective, however, as in the past, is to approve tests that have adequate statistical power to detect real contamination of ground water, and not to enforce minimum false positive rates.

This emphasis is evident in §264.98(g)(6) for detection monitoring and §264.99(i) for compliance monitoring. Both of these provisions allow the owner or operator to demonstrate that the statistically significant difference between background and compliance point wells or between compliance point wells and the Ground-Water Protection Standard is an artifact caused by an

error in sampling, analysis, statistical evaluation, or natural variation in ground-water chemistry. To make the demonstration that the statistically significant difference was caused by an error in sampling, analysis, or statistical evaluation, re-testing procedures that have been approved by the Regional Administrator can be written into the facility permit, provided their statistical power is comparable to the EPA Reference Power Curve given below.

For large monitoring networks, it is almost impossible to maintain a low network-wide overall false positive rate if the Type I errors for individual comparisons must be kept above 1%. As will be seen, some alternative testing strategies can achieve a low network-wide false positive rate while maintaining adequate power to detect contamination. EPA therefore recommends hat instead of the 1% criterion for individual comparisons, the overall network-wide false positive rate (across all wells and constituents) of any alternative testing strategy should be kept to approximately 5% for each monitoring or evaluation period, while maintaining statistical power comparable to the procedure below.

The other goal of any testing strategy should be to maintain adequate statistical power for detecting contamination. Technically, power refers to the probability that a statistical testing procedure will register and identify evidence of contamination when it exists. However, power is typically defined with respect to a single comparison, not a network of comparisons. Since some testing procedures may identify contamination more readily when several wells in the network are contaminated as opposed to just one or two, it is suggested that all testing strategies be compared on the following more stringent, but common, basis. Let the effective power of a testing procedure be defined as the probability of detecting contamination in the monitoring network when one and only one well is contaminated with a single constituent. Note that the effective power is a conservative measure of how a testing regimen will perform over the network, because the test must uncover one contaminated well among many clean ones (i.e., like "finding a needle in a haystack").

To establish a recommended standard for the statistical power of a testing strategy, it must be understood that the power is not single number, but rather a function of the level of contamination actually present. For most tests, the higher the level of contamination, the higher the statistical power; likewise, the lower the contamination level, the lower the power. As such, when increasingly contaminated ground water passes a particular well, it becomes easier for the statistical test to distinguish background levels from the contaminated ground water; consequently, the power is an increasing function of the contamination level.

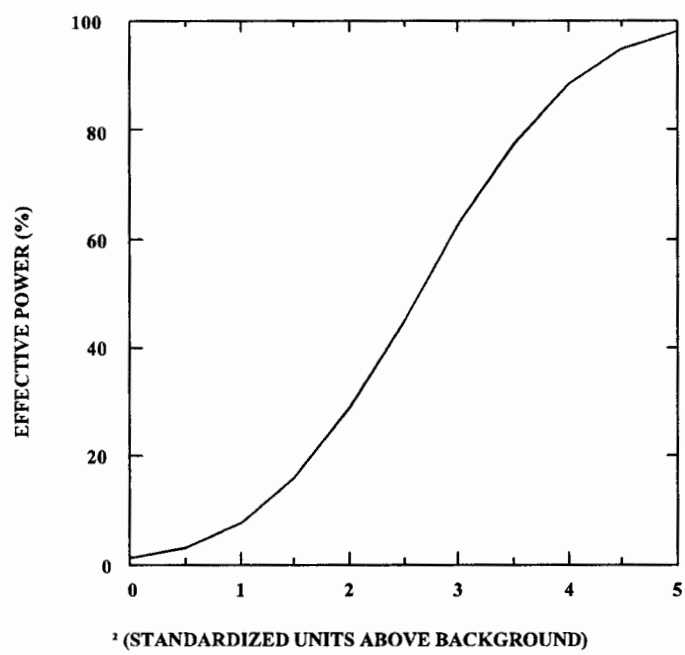
Perhaps the best way to describe the power function associated with a particular testing procedure is via a graph, such as the example below of the power of a standard Normal-based upper Prediction limit with 99% confidence. The power in percent is plotted along the y-axis against the standardized mean level of contamination along the x-axis. The standardized contamination levels are in units of standard deviations above the baseline (estimated from background data), allowing different power curves to be compared across indicator parameters, wells, and so forth. The standardized units, Δ, may be computed as

$$\Delta = \frac{\text{(Mean Contamination Level)} - \text{(Mean Background Level)}}{\text{(SD of Background Data)}}$$

In some situations, the probability that contamination will be detected by a particular testing procedure may be difficult if not impossible to derive analytically and will have to be simulated on a computer. In these cases, the power is typically estimated by generating Normally-distributed random values at different mean levels and repeatedly simulating the test procedure. With enough repetitions a reliable power curve can be plotted (e.g., see figure below).

EPA REFERENCE POWER CURVE

(16 Background Samples)



Notice that the power at $\Delta=0$ represents the false positive rate of the test, because at that point no contamination is actually present and the curve is indicating how often contamination will be "detected" anyway. As long as the power at $\Delta=0$ is approximately 5% (except for tests on an individual constituent at an individual well where the false positive rate should approximate 1%) and the rest of the power curve is acceptably high, the testing strategy should be adequately comparable to EPA standards.

To determine an acceptable power curve for comparison to alternative testing strategies, the following EPA Reference Power Curve is suggested. For a given and fixed number of background measurements, and based on Normally-distributed data from a single downgradient well generated at various mean levels above background, the EPA Reference Power Curve will represent the power associated with a 99% confidence upper prediction limit on the next single future sample from the well (see figure above for n=16).

Since the power of a test depends on several factors, including the background sample size, the type of test, and the number of comparisons, a different EPA Reference Power Curve will be associated with each distinct number of background samples. Power curves of alternative tests should only be compared to the EPA Reference Power Curve using a comparable number of background measurements. If the power of the alternative test is at least as high as the EPA reference, while maintaining an approximate 5% overall false positive rate, the alternative procedure should be acceptable.

With respect to power curves, keep in mind three important considerations: 1) the power of any testing method can be increased merely by relaxing the false positive rate requirement, letting α become larger than 5%. This is why an approximate 5% alpha level is suggested as the standard guidance, to ensure fair power comparisons among competing tests and to limit the overall network-wide false positive rate. 2) The simulation of alternative testing methods should incorporate every aspect of the procedure, from initial screens of the data to final decisions concerning the presence of contamination. This is especially applicable to strategies that involve some form of retesting at potentially contaminated wells. 3) When the testing strategy incorporates multiple comparisons, it is crucial that the power be gauged by simulating contamination in one and only one indicator parameter at a single well (i.e., by measuring the effective power). As noted earlier, EPA recommends that power be defined conservatively, forcing any test procedure to find "the needle in the haystack."

5.2 POSSIBLE STRATEGIES

5.2.1 Parametric and Non-parametric ANOVA

As described in the Interim Final Guidance, ANOVA procedures (either the parametric method or the Kruskal-Wallis test) allow multiple downgradient wells (but not multiple constituents) to be combined into a single statistical test, thus enabling the network-wide false positive rate for any single constituent to be kept at 5% regardless of the size of the network. The ANOVA method also maintains decent power for detecting real contamination, though only for small to moderately-sized networks. In large networks, even the parametric ANOVA has a difficult time finding the "needle in a haystack." The reason for this is that the ANOVA F-test combines all downgradient wells simultaneously, so that "clean" wells are mixed together with the single contaminated well, potentially masking the test's ability to detect the source of contamination.

Because of these characteristics, the ANOVA procedure may have poorer power for detecting a narrow plume of contamination which affects only one or two wells in a much larger network (say 20 or more comparisons). Another drawback is that a significant ANOVA test result will not indicate which well or wells is potentially contaminated without further post-hoc testing. Furthermore, the power of the ANOVA procedure depends significantly on having at least 3 to 4 samples per well available for testing. Since the samples must be statistically independent, collection of 3 or more samples at a given well may necessitate a several-month wait if the natural ground-water velocity at that well is low. In this case, it may be tempting to look for other strategies (e.g., Tolerance or Prediction intervals) that allow statistical testing of each new ground water sample as it is collected and analyzed. Finally, since the simple one-way ANOVA procedure outlined in the Interim Final Guidance is not designed to test multiple constituents simultaneously, the overall false positive rate will be approximately 5% <u>per constituent</u>, leading to a potentially high overall network-wide false positive rate (across wells and constituents) if many constituents need to be tested.

5.2.2 Retesting with Parametric Intervals

One strategy alternative to ANOVA is a modification of approaches suggested by Gibbons (1991a) and Davis and McNichols (1987). The basic idea is to adopt a two-phase testing strategy. First, new samples from each well in the network are compared, for each designated constituent parameter, against an upper Tolerance limit with pre-specified average coverage

(Note that the upper Tolerance limit will be different for each constituent). Since some constituents at some wells in a large network would be expected to fail the Tolerance limit even in the absence of contamination, each well that triggers the Tolerance limit is resampled and only those constituents that "triggered" the limit are retested via an upper Prediction limit (again differing by constituent). If one or more resamples fails the upper Prediction limit, the specific constituent at that well failing the test is deemed to have a concentration level significantly greater than background. The overall strategy is effective for large networks of comparisons (e.g., 100 or more comparisons), but also flexible enough to accommodate smaller networks.

To design and implement an appropriate pair of Tolerance and Prediction intervals, one must know the number of background samples available and the number of comparisons in the network. Since parametric intervals are used, it is assumed that the background data are either Normal or can be transformed to an approximate Normal distribution. The tricky part is to choose an average coverage for the Tolerance interval and confidence level for the Prediction interval such that the twin goals are met of keeping the overall false positive rate to approximately 5% and maintaining adequate statistical power.

To derive the overall false positive rate for this retesting strategy, assume that when no contamination is present each constituent and well in the network behaves independently of other constituents and wells. Then if A_i denotes the event that well i is triggered falsely at some stage of the testing, the overall false positive rate across m such comparisons can be written as

$$\text{total } \alpha = \Pr\{A_1 \text{ or } A_2 \text{ or } \ldots \text{ or } A_i \text{ or } \ldots \text{ or } A_m\} = 1 - \prod_{i=1}^{m} \Pr\{\overline{A}_i\}$$

where $\overline{A}_i$ denotes the complement of event A_i. Since $P\{\overline{A}_i\}$ is the probability of <u>not</u> registering a false trigger at uncontaminated well i, it may be written as

$$\Pr\{\overline{A}_i\} = \Pr\{X_i \le TL\} + \Pr\{X_i > TL\} \times \Pr\{Y_i \le PL \mid X_i > TL\}$$

where X_i represents the original sample at well i, Y_i represents the concentrations of one or more resamples at well i, TL and PL denote the upper Tolerance and Prediction limits respectively, and the right-most probability is the conditional event that all resample concentrations fall below the Prediction limit when the initial sample fails the Tolerance limit.

Letting $x=\Pr\{X_i \le TL\}$ and $y=\Pr\{Y_i \le PL \mid X_i > TL\}$, the overall false positive rate across m constituent-well combinations can be expressed as

$$\text{total } \alpha = 1 - [x + (1-x)\cdot y]^m$$

As noted by Guttman (1970), the probability that any random sample will fall below the upper Tolerance limit (i.e., quantity x above) is equal to the expected or average coverage of the Tolerance interval. If the Tolerance interval has been constructed to have average coverage of 95%, x=0.95. Then given a predetermined value for x, a fixed number of comparisons m, and a desired overall false positive rate α, we can solve for the conditional probability y as follows:

$$y = \frac{\sqrt[m]{1-\alpha} - x}{1-x}$$

If the conditional probability y were equal to the probability that the resample(s) for the ith constituent-well combination falls below the upper Prediction limit, one could fix α at, say, 5%, and construct the Prediction interval to have confidence level y. In that way, one could guarantee an expected network-wide false positive rate of 5%. Unfortunately, whether or not one or more resamples falls below the Prediction limit depends partly on whether the initial sample for that comparison eclipsed the Tolerance limit. This is because the same background data are used to construct both the Tolerance limit and the Prediction limit, creating a statistical dependence between the tests.

The exact relationship between the conditional probability y and the unconditional probability $\Pr\{Y_i \le PL\}$ is not known; however, simulations of the testing strategy suggest that when the confidence level for the Prediction interval is equated to the above solution for y, the overall network-wide false positive rate turns out to be higher than 5%. How much higher depends on the number of background samples and also the number of downgradient comparisons. Even with a choice of y that guarantees an expected facility-wide false positive rate of 5%, the power characteristics of the resulting testing strategy are not necessarily equivalent to the EPA Reference Power Curve, again depending on the number of background samples and the number of monitoring well-constituent combinations in the network.

In practice, to meet the selection criteria of 1) establishing an overall false positive rate of approximately 5% and 2) maintaining adequate statistical power, the confidence level chosen for the upper Prediction limit should be somewhat higher than the solution y to the preceding

equation. The table below provides recommended choices of expected coverage and confidence levels for the Tolerance interval-Prediction interval pair when using specific combinations of numbers of downgradient comparisons and background samples. In general, one should pick lower coverage Tolerance limits for smaller networks and higher coverage Tolerance limits for larger networks. That way (as can be seen in the table), the resulting Prediction limit confidence levels will be low enough to allow the construction of Prediction limits with decent statistical power.

PARAMETRIC RETESTING STRATEGIES				
# COMPARISONS	# BG SAMPLES	TOLERANCE COVERAGE (%)	PREDICTION LEVEL (%)	RATING
	8	95	90	**
	16	95	90	**
5	16	95	85	*
	24	95	85	**
	24	95	90	*
	8	95	98	**
20	16	95	97	**
	24	95	97	**
	16	98	97	**
	16	99	92	*
50	24	98	95	**
	24	99	90	**
	16	98	98	*
100	24	99	95	*
	24	98	98	*

Note: ** = strongly recommended
* = recommended

Only strategies that approximately met the selection criteria are listed in the table. It can be seen that some, but not all, of these strategies are strongly recommended. Those that are merely "recommended" failed in the simulations to fully meet one or both of the selection criteria. The performance of all the recommended strategies, however, should be adequate to correctly identify contamination while maintaining a modest facility-wide false positive rate.

Once a combination of coverage and confidence levels for the Tolerance-Prediction interval pair is selected, the statistical power of the testing strategy should be estimated in order to compare with the EPA Reference Power Curve (particularly if the testing scenario is different

from those computed in this Addendum). Simulation results have suggested that the above method for choosing a two-phase testing regimen can offer statistical power comparable to the EPA Reference for almost any sized monitoring network (see power curves in Appendix B).

Several examples of simulated power curves are presented in Appendix B. The range of downgradient wells tested is from 5 to 100 (note that the number of wells could actually represent the number of constituent-well combinations if testing multiple parameters), and each curve is based on either 8, 16, or 24 background samples. The y-axis of each graph measures the effective power of the testing strategy, i.e., the probability that contamination is detected when one and only one constituent at a single well has a mean concentration higher than background level. For each case, the EPA Reference Power Curve is compared to two different two-phase testing strategies. In the first case, wells that trigger the initial Tolerance limit are resampled once. This single resample is compared to a Prediction limit for the next future sample. In the second case, wells that trigger the Tolerance limit are resampled twice. Both resamples are compared to an upper Prediction limit for the next two future samples at that well.

The simulated power curves suggest two points. First, with an appropriate choice of coverage and prediction levels, the two-phase retesting strategies have comparable power to the EPA Reference Power Curve, while maintaining low overall network-wide false positive rates. Second, the power of the retesting strategy is slightly improved by the addition of a second resample at wells that fail the initial Tolerance limit, because the sample size is increased.

Overall, the two-phase testing strategy defined above--i.e., first screening the network of wells with a single upper Tolerance limit, and then applying an upper Prediction limit to resamples from wells which fail the Tolerance interval--appears to meet EPA's objectives of maintaining adequate statistical power for detecting contamination while limiting network-wide false positive rates to low levels. Furthermore, since each compliance well is compared against the interval limits separately, a narrow plume of contamination can be identified more efficiently than with an ANOVA procedure (e.g., no post-hoc testing is necessary to finger the guilty wells, and the two-phase interval testing method has more power against the "needle-in-a-haystack" contamination hypothesis).

5.2.3 Retesting with Non-parametric Intervals

When parametric intervals are not appropriate for the data at hand, either due to a large fraction of nondetects or a lack of fit to Normality or Lognormality, a network of individual

comparisons can be handled via retesting using non-parametric Prediction limits. The strategy is to establish a non-parametric prediction limit for each designated indicator parameter based on background samples that accounts for the number of well-constituent comparisons in the overall network.

In order to meet the twin goals of maintaining adequate statistical power and a low overall rate of false positives, a non-parametric strategy must involve some level of retesting at those wells which initially indicate possible contamination. Retesting can be accomplished by taking a specific number of additional, independent samples from each well in which a specific constituent triggers the initial test and then comparing these samples against the non-parametric prediction limit for that parameter.

Because more independent data is added to the overall testing procedure, retesting of additional samples, in general, enables one to make more powerful and more accurate determinations of possible contamination. Retesting does, however, involve a trade-off. Because the power of the test increases with the number of resamples, one must decide how quickly resamples can be collected to ensure 1) quick identification and confirmation of contamination and yet, 2) the statistical independence of successive resamples from any particular well. Do not forget that the performance of a non-parametric retesting strategy depends substantially on the independence of the data from each well.

Two basic approaches to non-parametric retesting have been suggested by Gibbons (1990 and 1991b). Both strategies define the upper Prediction limit for each designated parameter to be the maximum value of that constituent in the set of background data. Consequently, the background wells used to construct the limits must be uncontaminated. After the Prediction limits have been calculated, one sample is collected from each downgradient well in the network. If any sample constituent value is greater than its upper prediction limit, the initial test is "triggered" and one or more resamples must be collected at that downgradient well on the constituent for further testing.

At this point, the similarity between the two approaches ends. In his 1990 article, Gibbons computes the probability that at least one of m independent samples taken from each of k downgradient wells will be below (i.e., pass) the prediction limit. The m samples include both the initial sample and (m-1) resamples. Because retesting only occurs when the initial well sample fails the limit, a given well fails the overall test (initial comparison plus retests) only if all (m-1)

resamples are above the prediction limit. If any resample passes the prediction limit, that well is regarded as showing no significant evidence of contamination.

Initially, this first strategy may not appear to be adequately sensitive to mild contamination at a given downgradient well. For example, suppose two resamples are to be collected whenever the initial sample fails the upper prediction limit. If the initial sample is above the background maximum and one of the resamples is also above the prediction limit, the well can still be classified as "clean" if the other resample is below the prediction limit. Statistical power simulations (see Appendix B), however, suggest that this strategy will perform adequately under a number of monitoring scenarios. Still, EPA recognizes that a retesting strategy which might classify a well as "clean" when the initial sample and a resample both fail the upper Prediction limit could offer problematic implications for permit writers and enforcement personnel.

A more stringent approach was suggested by Gibbons in 1991. In that article (1991b), Gibbons computes, as "passing behavior," the probability that all but one of m samples taken from each of k wells pass the upper prediction limit. Under this definition, if the initial sample fails the upper Prediction limit, all (m-1) resamples must pass the limit in order for well to be classified as "clean" during that testing period. Consequently, if any single resample falls above the background maximum, that well is judged as showing significant evidence of contamination.

Either non-parametric retesting approach offers the advantage of being extremely easy to implement in field testing of a large downgradient well network. In practice, one has only to determine the maximum background sample to establish the upper prediction limit against which all other comparisons are made. Gibbons' 1991 retesting scheme offers the additional advantage of requiring less overall sampling at a given well to establish significant evidence of contamination. Why? If the testing procedure calls for, say, two resamples at any well that fails the initial prediction limit screen, retesting can end whenever either one of the two resamples falls above the prediction limit. That is, the well will be designated as potentially contaminated if the first resample fails the prediction limit even if the second resample has not yet been collected.

In both of his papers, Gibbons offers tables that can be used to compute the overall network-wide false positive rate, given the number of background samples, the number of downgradient comparisons, and the number of retests for each comparison. It is clear that there is less flexibility in adjusting a non-parametric as opposed to a parametric prediction limit to achieve a certain Type I error rate. In fact, if only a certain number of retests are feasible at any given well (e.g., in order to maintain independence of successive samples), the only recourse to maintain

a low false positive rate is to collect a larger number of background samples. In this way, the inability to make parametric assumptions about the data illustrates why non-parametric tests are on the whole less efficient and less powerful than their parametric counterparts.

Unfortunately, the power of these non-parametric retesting strategies is not explored in detail by Gibbons. To compare the power of both Gibbons' strategies against the EPA Reference Power Curve, Normally distributed data were simulated for several combinations of numbers of background samples and downgradient wells (again, if multiple constituents are being tested, the number of wells in the simulations may be regarded as the number of constituent-well combinations). Up to three resamples were allowed in the simulations for comparative purposes. EPA recognizes, however, that it will be feasible in general to collect only one or two independent resamples from any given well. Power curves representing the results of these simulations are given in Appendix B. For each scenario, the EPA Reference Power Curve is compared with the simulated powers of six different testing strategies. These strategies include collection of no resamples, one resample, two resamples under Gibbons' 1990 approach (designated as A on the curves) and his 1991 approach (labelled as B), and three resamples (under approaches A and B). Under the one resample strategy, a potentially contaminated compliance well is designated as "clean" if the resample passes the retest and "contaminated" otherwise.

The following table lists the best-performing strategies under each scenario. As with the use of parametric intervals for retesting, the criteria for selecting the best-performing strategies required 1) an approximate 5% facility-wide false positive rate and 2) power equivalent to or better than the EPA Reference Power Curve. Because Normal data were used in these power simulations, more realistically skewed data would likely result in greater advantages for the non-parametric retesting strategies over the EPA Reference test.

Examination of the table and the power curves in Appendix B shows that the number of background samples has an important effect on the recommended testing strategy. For instance, with 8 background samples in a network of at least 20 wells, the best performing strategies all involve collection of 3 resamples per "triggered" compliance well (EPA regards such a strategy as impractical for permitting and enforcement purposes at most RCRA facilities). It tends to be true that as the number of available background samples grows, fewer resamples are needed from each potentially contaminated compliance well to maintain adequate power. If, as is expected, the number of feasible, independent retests is limited, a facility operator may have to collect additional background measurements in order to establish an adequate retesting strategy.

NON-PARAMETRIC RETESTING STRATEGIES				
# WELLS	# BG SAMPLES	STRATEGY	REFERENCE	RATING
	8	1 Resample		*
5	8	2 Resamples (A)	Gibbons, 1990	**
	16	1 Resample		**
	16	2 Resamples (B)	Gibbons, 1991	**
	24	2 Resamples (B)	Gibbons, 1991	**
	8	2 Resamples (A)	Gibbons, 1990	*
	16	1 Resample		*
20	16	2 Resamples (A)	Gibbons, 1990	*
	24	1 Resample		**
	24	2 Resamples (B)	Gibbons, 1991	*
	32	1 Resample		*
	32	2 Resamples (B)	Gibbons, 1991	**
	16	2 Resamples (A)	Gibbons, 1990	**
50	24	1 Resample		*
	24	2 Resamples (A)	Gibbons, 1990	*
	32	1 Resample		**
100	16	2 Resamples (A)	Gibbons, 1990	**
	24	2 Resamples (A)	Gibbons, 1990	*
	32	1 Resample		*

Note: ** = very good performance * = good performance

6. OTHER TOPICS

6.1 CONTROL CHARTS

Control Charts are an alternative to Prediction limits for performing either intrawell comparisons or comparisons to historically monitored background wells during detection monitoring. Since the baseline parameters for a Control Chart are estimated from historical data, this method is only appropriate for initially uncontaminated compliance wells. The main advantage of a Control Chart over a Prediction limit is that a Control Chart allows data from a well to be viewed graphically over time. Trends and changes in the concentration levels can be seen easily, because all sample data is consecutively plotted on the chart as it is collected, giving the data analyst an historical overview of the pattern of contamination. Prediction limits allow only point-in-time comparisons between the most recent data and past information, making long-term trends difficult to identify.

More generally, intrawell comparison methods eliminate the need to worry about spatial variability between wells in different locations. Whenever background data is compared to compliance point measurements, there is a risk that any statistically significant difference in concentration levels is due to spatial and/or hydrogeological differences between the wells rather than contamination at the facility. Because intrawell comparisons involve but a single well, significant changes in the level of contamination cannot be attributed to spatial differences between wells, regardless of whether the method used is a Prediction limit or Control Chart.

Of course, past observations can be used as baseline data in an intrawell comparison only if the well is known to be uncontaminated. Otherwise, the comparison between baseline data and newly collected samples may negate the goal in detection monitoring of identifying evidence of contamination. Furthermore, without specialized modification, Control Charts do not efficiently handle truncated data sets (i.e., those with a significant fraction of nondetects), making them appropriate only for those constituents with a high frequency of occurrence in monitoring wells. Control Charts tend to be most useful, therefore, for inorganic parameters (e.g., some metals and geochemical monitoring parameters) that occur naturally in the ground water.

The steps to construct a Control Chart can be found on pp. 7-3 to 7-10 of the Interim Final Guidance. The way a Control Chart works is as follows. Initial sample data is collected (from the specific compliance well in an intrawell comparison or from background wells in comparisons of compliance data with background) in order to establish baseline parameters for the chart, specifically, estimates of the well mean and well variance. These samples are meant to characterize the concentration levels of the uncontaminated well, before the onset of detection monitoring. Since the estimate of well variance is particularly important, it is recommended that at least 8 samples be collected (say, over a year's time) to estimate the baseline parameters. Note that none of these 8 or more samples is actually plotted on the chart.

As future samples are collected, the baseline parameters are used to standardize the data. At each sampling period, a standardized mean is computed using the formula below, where m represents the baseline mean concentration and s represents the baseline standard deviation.

$$Z_i = \sqrt{n_i}\,(\bar{x} - m) / s$$

A cumulative sum (CUSUM) for the ith period is also computed, using the formula $S_i = \max\{0, (Z_i-k)+S_{i-1}\}$, where Z_i is the standardized mean for that period and k represents a pre-chosen Control Chart parameter.

Once the data have been standardized and plotted, a Control Chart is declared out-of-control if the sample concentrations become too large when compared to the baseline parameters. An out-of-control situation is indicated on the Control Chart when either the standardized means or CUSUMs cross one of two pre-determined threshold values. These thresholds are based on the rationale that if the well remains uncontaminated, new sample values standardized by the original baseline parameters should not deviate substantially from the baseline level. If contamination does occur, the old baseline parameters will no longer accurately represent concentration levels at the well and, hence, the standardized values should significantly deviate from the baseline levels on the Control Chart.

In the combined Shewhart-cumulative sum (CUSUM) Control Chart recommended by the Interim Final Guidance (Section 7), the chart is declared out-of-control in one of two ways. First, the standardized means (Z_i) computed at each sampling period may cross the Shewhart control limit (SCL). Such a change signifies a rapid increase in well concentration levels among the most recent sample data. Second, the cumulative sum (CUSUM) of the standardized means may become too large, crossing the "decision internal value" (h). Crossing the h threshold can mean either a sudden rise in concentration levels or a gradual increase over a longer span of time. A gradual increase or trend is particularly indicated if the CUSUM crosses its threshold but the standardized mean Z_i does not. The reason for this is that several consecutive small increases in Z_i will not trigger the SCL threshold, but may trigger the CUSUM threshold. As such, the Control Chart can indicate the onset of either sudden or gradual contamination at the compliance point.

As with other statistical methods, Control Charts are based on certain assumptions about the sample data. The first is that the data at an uncontaminated well (i.e., a well process that is "in control") are Normally distributed. Since estimates of the baseline parameters are made using initially collected data, these data should be tested for Normality using one of the goodness-of-fit techniques described earlier. Better yet, the logarithms of the data should be tested first, to see if a Lognormal model is appropriate for the concentration data. If the Lognormal model is not rejected, the Control Chart should be constructed solely on the basis of logged data.

The methodology for Control Charts also assumes that the sample data are independently distributed from a statistical standpoint. In fact, these charts can easily give misleading results if the consecutive sample data are not independent. For this reason, it is important to design a sampling plan so that distinct volumes of water are analyzed each sampling period and that

duplicate sample analyses are not treated are independent observations when constructing the Control Chart.

The final assumption is that the baseline parameters at the well reflect current background concentration levels. Some long-term fluctuation in background levels may be possible even though contamination has not occurred at a given well. Because of this possibility, if a Control Chart remains "in control" for a long period of time, the baseline parameters should be updated to include more recent observations as background data. After all, the original baseline parameters will often be based only on the first year's data. Much better estimates of the true background mean and variance can be obtained by including more data at a later time.

To update older background data with more recent samples, a two-sample t-test can be run to compare the older concentration levels with the concentrations of the proposed update samples. If the t-test does not show a significant difference at the 5 percent significance level, proceed to re-estimate the baseline parameters by including more recent data. If the t-test does show a significant difference, the newer data should not be characterized as background unless some specific factor can be pinpointed explaining why background levels on the site have naturally changed.

EXAMPLE 18

Construct a control chart for the 8 months of data collected below.

μ=27 ppb
σ=25 ppb

	Nickel Concentration (ppb)	
Month	Sample 1	Sample 2
1	15.3	22.6
2	41.1	27.8
3	17.5	18.1
4	15.7	31.5
5	37.2	32.4
6	25.1	32.5
7	19.9	27.5
8	99.3	64.2

SOLUTION

Step 1. The three parameters necessary to construct a combined Shewhart-CUSUM chart are h=5, k=1, and SCL=4.5 in units of standard deviation (SD).

Step 2. List the sampling periods and monthly means, as in the following table.

Month	T_i	Mean (ppb)	Z_i	Z_i - k	S_i
1	1	19.0	-0.45	-1.45	0.00
2	2	34.5	0.42	-0.58	0.00
3	3	17.8	-0.52	-1.52	0.00
4	4	23.6	-0.19	-1.19	0.00
5	5	34.8	0.44	-0.56	0.00
6	6	28.8	0.10	-0.90	0.00
7	7	23.7	-0.19	-1.19	0.00
8	8	81.8	3.10	2.10	2.10

Step 3. Compute the standardized means Z_i and the quantities S_i. List in the table above. Each S_i is computed for consecutive months using the formula on p. 7-8 of the EPA guidance document.

$$S_1 = \max\{0, -1.45 + 0\} = 0.00$$

$$S_2 = \max\{0, -0.58 + 0\} = 0.00$$

$$S_3 = \max\{0, -1.52 + 0\} = 0.00$$

$$S_4 = \max\{0, -1.19 + 0\} = 0.00$$

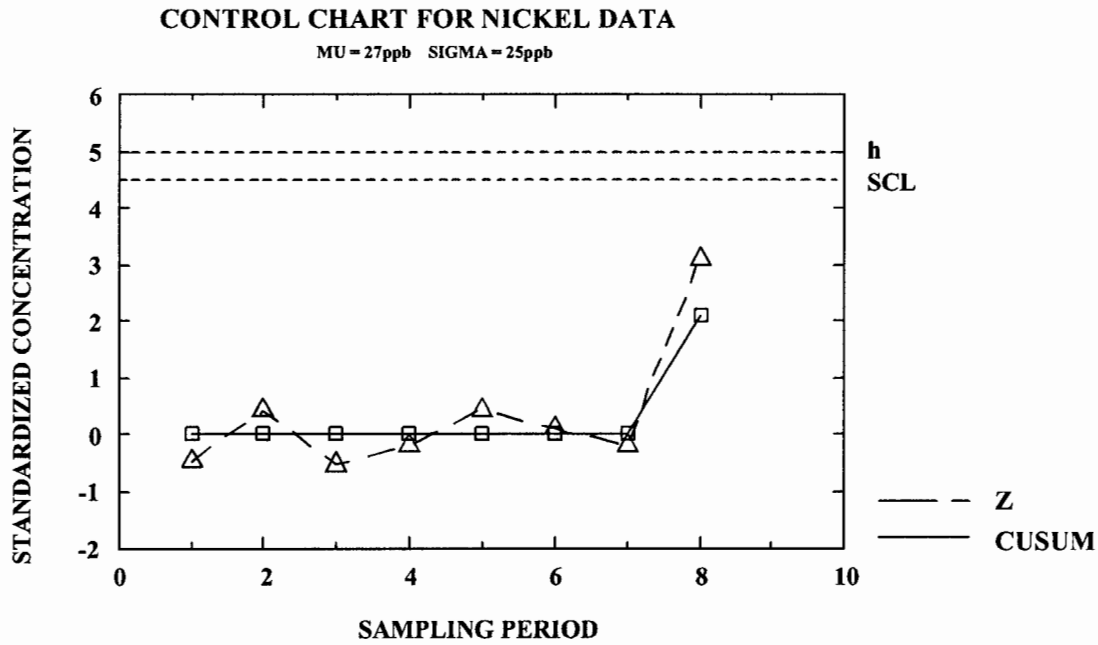
$$S_5 = \max\{0, -0.56 + 0\} = 0.00$$

$$S_6 = \max\{0, -0.90 + 0\} = 0.00$$

$$S_7 = \max\{0, -1.19 + 0\} = 0.00$$

$$S_8 = \max\{0, 2.10 + 0\} = 2.10$$

Step 4. Plot the control chart as given below. The combined chart indicates that there is no evidence of contamination at the monitoring facility because neither the standardized mean nor the CUSUM statistic exceeds the Shewhart control limits for the months examined.



Note: In the above Control Chart, the CUSUMs are compared to threshold h, while the standardized means (Z) are compared to the SCL threshold.

6.2 OUTLIER TESTING

Formal testing for outliers should be done only if an observation seems particularly high (by orders of magnitude) compared to the rest of the data set. If a sample value is suspect, one should run the outlier test described on pp. 8-11 to 8-14 of the EPA guidance document. It should be cautioned, however, that this outlier test assumes that the rest of the data values, except for the suspect observation, are Normally distributed (Barnett and Lewis, 1978). Since Lognormally distributed measurements often contain one or more values that appear high relative to the rest, it is recommended that the outlier test be run on the logarithms of the data instead of the original observations. That way, one can avoid classifying a high Lognormal measurement as an outlier just because the test assumptions were violated.

If the test designates an observation as a statistical outlier, the sample should not be treated as such until a specific reason for the abnormal measurement can be determined. Valid reasons may, for example, include contaminated sampling equipment, laboratory contamination of the

sample, or errors in transcription of the data values. Once a specific reason is documented, the sample should be excluded from any further statistical analysis. If a plausible reason cannot be found, the sample should be treated as a true but extreme value, not to be excluded from further analysis.

EXAMPLE 19

The table below contains data from five wells measured over a 4-month period. The value 7066 is found in the second month at well 3. Determine whether there is statistical evidence that this observation is an outlier.

Carbon Tetrachloride Concentration (ppb)				
Well 1	Well 2	Well 3	Well 4	Well 5
1.69	302	16.2	199	275
3.25	35.1	7066	41.6	6.5
7.3	15.6	350	75.4	59.7
12.1	13.7	70.14	57.9	68.4

SOLUTION

Step 1. Take logarithms of each observation. Then order and list the logged concentrations.

Order	Concentration (ppb)	Logged Concentration
1	1.69	0.525
2	3.25	1.179
3	6.5	1.872
4	7.3	1.988
5	12.1	2.493
6	13.7	2.617
7	15.6	2.747
8	16.2	2.785
9	35.1	3.558
10	41.6	3.728
11	57.9	4.059
12	59.7	4.089
13	68.4	4.225
14	70.1	4.250
15	75.4	4.323
16	199	5.293
17	275	5.617
18	302	5.710
19	350	5.878
20	7066	8.863

Step 2. Calculate the mean and SD of all the logged measurements. In this case, the mean and SD are 3.789 and 1.916, respectively.

Step 3. Calculate the outlier test statistic T_{20} as

$$T_{20} = \frac{X_{(20)} - \overline{X}}{SD} = \frac{8.863 - 3.789}{1.916} = 2.648.$$

Step 4. Compare the observed statistic T_{20} with the critical value of 2.557 for a sample size n=20 and a significance level of 5 percent (taken from Table 8 on p. B-12 of the Interim Final Guidance). Since the observed value T_{20}=2.648 exceeds the critical value, there is significant evidence that the largest observation is a statistical outlier. Before excluding this value from further analysis, a valid explanation for this unusually high value should be found. Otherwise, treat the outlier as an extreme but valid concentration measurement.

TABLE A-1.

COEFFICIENTS $\{A_{N-I+1}\}$ FOR W TEST OF NORMALITY, FOR N=2(1)50

i/n	2	3	4	5	6	7	8	9	10
1	0.7071	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739
2	----	.0000	.1677	.2413	.2806	.3031	.3164	.3244	.3291
3	----	----	----	.0000	.0875	.1401	.1743	.1976	.2141
4	----	----	----	----	----	.0000	.0561	.0947	.1224
5	----	----	----	----	----	----	----	.0000	.0399

i/n	11	12	13	14	15	16	17	18	19	20
1	0.5601	0.5475	0.5359	0.5251	0.5150	0.5056	0.4968	0.4886	0.4808	0.4734
2	.3315	.3325	.3325	.3318	.3306	.3290	.3273	.3253	.3232	.3211
3	.2260	.2347	.2412	.2460	.2495	.2521	.2540	.2553	.2561	.2565
4	.1429	.1586	.1707	.1802	.1878	.1939	.1988	.2027	.2059	.2085
5	.0695	.0922	.1099	.1240	.1353	.1447	.1524	.1587	.1641	.1686
6	0.0000	0.0303	0.0539	0.0727	0.0880	0.1005	0.1109	0.1197	0.1271	0.1334
7	----	----	.0000	.0240	.0433	.0593	.0725	.0837	.0932	.1013
8	----	----	----	----	.0000	.0196	.0359	.0496	.0612	.0711
9	----	----	----	----	----	----	.0000	.0163	.0303	.0422
10	----	----	----	----	----	----	----	----	.0000	.0140

i/n	21	22	23	24	25	26	27	28	29	30
1	0.4643	0.4590	0.4542	0.4493	0.4450	0.4407	0.4366	0.4328	0.4291	0.4254
2	.3185	.3156	.3126	.3098	.3069	.3043	.3018	.2992	.2968	.2944
3	.2578	.2571	.2563	.2554	.2543	.2533	.2522	.2510	.2499	.2487
4	.2119	.2131	.2139	.2145	.2148	.2151	.2152	.2151	.2150	.2148
5	.1736	.1764	.1787	.1807	.1822	.1836	.1848	.1857	.1864	.1870
6	0.1399	0.1443	0.1480	0.1512	0.1539	0.1563	0.1584	0.1601	0.1616	0.1630
7	.1092	.1150	.1201	.1245	.1283	.1316	.1346	.1372	.1395	.1415
8	.0804	.0878	.0941	.0997	.1046	.1089	.1128	.1162	.1192	.1219
9	.0530	.0618	.0696	.0764	.0823	.0876	.0923	.0965	.1002	.1036
10	.0263	.0368	.0459	.0539	.0610	.0672	.0728	.0778	.0822	.0862
11	0.0000	0.0122	0.0228	0.0321	0.0403	0.0476	0.0540	0.0598	0.0650	0.0697
12	----	----	.0000	.0107	.0200	.0284	.0358	.0424	.0483	.0537
13	----	----	----	----	.0000	.0094	.0178	.0253	.0320	.0381
14	----	----	----	----	----	----	.0000	.0084	.0159	.0227
15	----	----	----	----	----	----	----	----	.0000	.0076

i/n	31	32	33	34	35	36	37	38	39	40
1	0.4220	0.4188	0.4156	0.4127	0.4096	0.4068	0.4040	0.4015	0.3989	0.3964
2	.2921	.2898	.2876	.2854	.2834	.2813	.2794	.2774	.2755	.2737
3	.2475	.2463	.2451	.2439	.2427	.2415	.2403	.2391	.2380	.2368
4	.2145	.2141	.2137	.2132	.2127	.2121	.2116	.2110	.2104	.2098
5	.1874	.1878	.1880	.1882	.1883	.1883	.1883	.1881	.1880	.1878
6	0.1641	0.1651	0.1660	0.1667	0.1673	0.1678	0.1683	0.1686	0.1689	0.1691
7	.1433	.1449	.1463	.1475	.1487	.1496	.1503	.1513	.1520	.1526
8	.1243	.1265	.1284	.1301	.1317	.1331	.1344	.1356	.1366	.1376
9	.1066	.1093	.1118	.1140	.1160	.1179	.1196	.1211	.1225	.1237
10	.0899	.0931	.0961	.0988	.1013	.1036	.1056	.1075	.1092	.1108

TABLE A-1. (CONTINUED)

COEFFICIENTS $\{A_{N-I+1}\}$ FOR W TEST OF NORMALITY, FOR N=2(1)50

i/n	31	32	33	34	35	36	37	38	39	40
11	0.0739	0.0777	0.0812	0.0844	0.0873	0.0900	0.0924	0.0947	0.0967	0.0986
12	.0585	.0629	.0669	.0706	.0739	.0770	.0798	.0824	.0848	.0870
13	.0435	.0485	.0530	.0572	.0610	.0645	.0677	.0706	.0733	.0759
14	.0289	.0344	.0395	.0441	.0484	.0523	.0559	.0592	.0622	.0651
15	.0144	.0206	.0262	.0314	.0361	.0404	.0444	.0481	.0515	.0546
16	0.0000	0.0068	0.0131	0.0187	0.0239	0.0287	0.0331	0.0372	0.0409	0.0444
17	----	----	.0000	.0062	.0119	.0172	.0220	.0264	.0305	.0343
18	----	----	----	----	.0000	.0057	.0110	.0158	.0203	.0244
19	----	----	----	----	----	----	.0000	.0053	.0101	.0146
20	----	----	----	----	----	----	----	----	.0000	.0049

i/n	41	42	43	44	45	46	47	48	49	50
1	0.3940	0.3917	0.3894	0.3872	0.3850	0.3830	0.3808	0.3789	0.3770	0.3751
2	.2719	.2701	.2684	.2667	.2651	.2635	.2620	.2604	.2589	.2574
3	.2357	.2345	.2334	.2323	.2313	.2302	.2291	.2281	.2271	.2260
4	.2091	.2085	.2078	.2072	.2065	.2058	.2052	.2045	.2038	.2032
5	.1876	.1874	.1871	.1868	.1865	.1862	.1859	.1855	.1851	.1847
6	0.1693	0.1694	0.1695	0.1695	0.1695	0.1695	0.1695	0.1693	0.1692	0.1691
7	.1531	.1535	.1539	.1542	.1545	.1548	.1550	.1551	.1553	.1554
8	.1384	.1392	.1398	.1405	.1410	.1415	.1420	.1423	.1427	.1430
9	.1249	.1259	.1269	.1278	.1286	.1293	.1300	.1306	.1312	.1317
10	.1123	.1136	.1149	.1160	.1170	.1180	.1189	.1197	.1205	.1212
11	0.1004	0.1020	0.1035	0.1049	0.1062	0.1073	0.1085	0.1095	0.1105	0.1113
12	.0891	.0909	.0927	.0943	.0959	.0972	.0986	.0998	.1010	.1020
13	.0782	.0804	.0824	.0842	.0860	.0876	.0892	.0906	.0919	.0932
14	.0677	.0701	.0724	.0745	.0775	.0785	.0801	.0817	.0832	.0846
15	.0575	.0602	.0628	.0651	.0673	.0694	.0713	.0731	.0748	.0764
16	0.0476	0.0506	0.0534	0.0560	0.0584	0.0607	0.0628	0.0648	0.0667	0.0685
17	.0379	.0411	.0442	.0471	.0497	.0522	.0546	.0568	.0588	.0608
18	.0283	.0318	.0352	.0383	.0412	.0439	.0465	.0489	.0511	.0532
19	.0188	.0227	.0263	.0296	.0328	.0357	.0385	.0411	.0436	.0459
20	.0094	.0136	.0175	.0211	.0245	.0277	.0307	.0335	.0361	.0386
21	0.0000	0.0045	0.0087	0.0126	0.0163	0.0197	0.0229	0.0259	0.0288	0.0314
22	----	----	.0000	.0042	.0081	.0118	.0153	.0185	.0215	.0244
23	----	----	----	----	.0000	.0039	.0076	.0111	.0143	.0174
24	----	----	----	----	----	----	.0000	.0037	.0071	.0104
25	----	----	----	----	----	----	----	----	.0000	.0035

TABLE A-2.

PERCENTAGE POINTS OF THE W TEST FOR N=3(1)50

n	0.01	0.05
3	0.753	0.767
4	.687	.748
5	.686	.762
6	0.713	0.788
7	.730	.803
8	.749	.818
9	.764	.829
10	.781	.842
11	0.792	0.850
12	.805	.859
13	.814	.866
14	.825	.874
15	.835	.881
16	0.844	0.887
17	.851	.892
18	.858	.897
19	.863	.901
20	.868	.905
21	0.873	0.908
22	.878	.911
23	.881	.914
24	.884	.916
25	.888	.918
26	0.891	0.920
27	.894	.923
28	.896	.924
29	.898	.926
30	.900	.927
31	0.902	0.929
32	.904	.930
33	.906	.931
34	.908	.933
35	.910	.934

TABLE A-2. (CONTINUED)

PERCENTAGE POINTS OF THE W TEST FOR N=3(1)50

n	0.01	0.05
36	0.912	0.935
37	.914	.936
38	.916	.938
39	.917	.939
40	.919	.940
41	0.920	0.941
42	.922	.942
43	.923	.943
44	.924	.944
45	.926	.945
46	0.927	0.945
47	.928	.946
48	.929	.947
49	.929	.947
50	.930	.947

TABLE A-3.

PERCENTAGE POINTS OF THE W′ TEST FOR N≥35

n	.01	.05
35	0.919	0.943
50	.935	.953
51	0.935	0.954
53	.938	.957
55	.940	.958
57	.944	.961
59	.945	.962
61	0.947	0.963
63	.947	.964
65	.948	.965
67	.950	.966
69	.951	.966
71	0.953	0.967
73	.956	.968
75	.956	.969
77	.957	.969
79	.957	.970
81	0.958	0.970
83	.960	.971
85	.961	.972
87	.961	.972
89	.961	.972
91	0.962	0.973
93	.963	.973
95	.965	.974
97	.965	.975
99	.967	.976

TABLE A-4.

PERCENT POINTS OF THE NORMAL PROBABILITY PLOT CORRELETION COEFFICIENT FOR N=3(1)50(5)100

n	.01	.025	.05
3	.869	.872	.879
4	.822	.845	.868
5	.822	.855	.879
6	.835	.868	.890
7	.847	.876	.899
8	.859	.886	.905
9	.868	.893	.912
10	.876	.900	.917
11	.883	.906	.922
12	.889	.912	.926
13	.895	.917	.931
14	.901	.921	.934
15	.907	.925	.937
16	.912	.928	.940
17	.912	.931	.942
18	.919	.934	.945
19	.923	.937	.947
20	.925	.939	.950
21	.928	.942	.952
22	.930	.944	.954
23	.933	.947	.955
24	.936	.949	.957
25	.937	.950	.958
26	.939	.952	.959
27	.941	.953	.960
28	.943	.955	.962
29	.945	.956	.962
30	.947	.957	.964
31	.948	.958	.965
32	.949	.959	.966
33	.950	.960	.967
34	.951	.960	.967
35	.952	.961	.968
36	.953	.962	.968
37	.955	.962	.969
38	.956	.964	.970
39	.957	.965	.971
40	.958	.966	.972

TABLE A-4. (CONTINUED)

PERCENT POINTS OF THE NORMAL PROBABILITY PLOT CORRELETION COEFFICIENT FOR N=3(1)50(5)100

n	.01	.025	.05
41	.958	.967	.973
42	.959	.967	.973
43	.959	.967	.973
44	.960	.968	.974
45	.961	.969	.974
46	.962	.969	.974
47	.963	.970	.975
48	.963	.970	.975
49	.964	.971	.977
50	.965	.972	.978
55	.967	.974	.980
60	.970	.976	.981
65	.972	.977	.982
70	.974	.978	.983
75	.975	.979	.984
80	.976	.980	.985
85	.977	.981	.985
90	.978	.982	.985
95	.979	.983	.986
100	.981	.984	.987

TABLE A-5.

VALUES OF LAMBDA FOR COHEN'S METHOD

	Percentage of Non-detects										
γ	.01	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
.01	.0102	.0530	.1111	.1747	.2443	.3205	.4043	.4967	.5989	.7128	.8403
.05	.0105	.0547	.1143	.1793	.2503	.3279	.4130	.5066	.6101	.7252	.8540
.10	.0110	.0566	.1180	.1848	.2574	.3366	.4233	.5184	.6234	.7400	.8703
.15	.0113	.0584	.1215	.1898	.2640	.3448	.4330	.5296	.6361	.7542	.8860
.20	.0116	.0600	.1247	.1946	.2703	.3525	.4422	.5403	.6483	.7678	.9012
.25	.0120	.0615	.1277	.1991	.2763	.3599	.4510	.5506	.6600	.7810	.9158
.30	.0122	.0630	.1306	.2034	.2819	.3670	.4595	.5604	.6713	.7937	.9300
.35	.0125	.0643	.1333	.2075	.2874	.3738	.4676	.5699	.6821	.8060	.9437
.40	.0128	.0657	.1360	.2114	.2926	.3803	.4755	.5791	.6927	.8179	.9570
.45	.0130	.0669	.1385	.2152	.2976	.3866	.4831	.5880	.7029	.8295	.9700
.50	.0133	.0681	.1409	.2188	.3025	.3928	.4904	.5967	.7129	.8408	.9826
.55	.0135	.0693	.1432	.2224	.3073	.3987	.4976	.6051	.7225	.8517	.9950
.60	.0137	.0704	.1455	.2258	.3118	.4045	.5046	.6133	.7320	.8625	1.0070
.65	.0140	.0715	.1477	.2291	.3163	.4101	.5114	.6213	.7412	.8729	1.0188
.70	.0142	.0726	.1499	.2323	.3206	.4156	.5180	.6291	.7502	.8832	1.0303
.75	.0144	.0736	.1520	.2355	.3249	.4209	.5245	.6367	.7590	.8932	1.0416
.80	.0146	.0747	.1540	.2386	.3290	.4261	.5308	.6441	.7676	.9031	1.0527
.85	.0148	.0756	.1560	.2416	.3331	.4312	.5370	.6515	.7761	.9127	1.0636
.90	.0150	.0766	.1579	.2445	.3370	.4362	.5430	.6586	.7844	.9222	1.0743
.95	.0152	.0775	.1598	.2474	.3409	.4411	.5490	.6656	.7925	.9314	1.0847
1.00	.0153	.0785	.1617	.2502	.3447	.4459	.5548	.6725	.8005	.9406	1.0951
1.05	.0155	.0794	.1635	.2530	.3484	.4506	.5605	.6793	.8084	.9496	1.1052
1.10	.0157	.0803	.1653	.2557	.3521	.4553	.5662	.6860	.8161	.9584	1.1152
1.15	.0159	.0811	.1671	.2584	.3557	.4598	.5717	.6925	.8237	.9671	1.1250
1.20	.0160	.0820	.1688	.2610	.3592	.4643	.5771	.6990	.8312	.9756	1.1347
1.25	.0162	.0828	.1705	.2636	.3627	.4687	.5825	.7053	.8385	.9841	1.1443
1.30	.0164	.0836	.1722	.2661	.3661	.4730	.5878	.7115	.8458	.9924	1.1537
1.35	.0165	.0845	.1738	.2686	.3695	.4773	.5930	.7177	.8529	1.0006	1.1629
1.40	.0167	.0853	.1754	.2710	.3728	.4815	.5981	.7238	.8600	1.0087	1.1721
1.45	.0168	.0860	.1770	.2735	.3761	.4856	.6031	.7298	.8670	1.0166	1.1812
1.50	.0170	.0868	.1786	.2758	.3793	.4897	.6081	.7357	.8738	1.0245	1.1901
1.55	.0171	.0876	.1801	.2782	.3825	.4938	.6130	.7415	.8806	1.0323	1.1989
1.60	.0173	.0883	.1817	.2805	.3856	.4977	.6179	.7472	.8873	1.0400	1.2076
1.65	.0174	.0891	.1832	.2828	.3887	.5017	.6227	.7529	.8939	1.0476	1.2162
1.70	.0176	.0898	.1846	.2851	.3918	.5055	.6274	.7585	.9005	1.0551	1.2248
1.75	.0177	.0905	.1861	.2873	.3948	.5094	.6321	.7641	.9069	1.0625	1.2332
1.80	.0179	.0913	.1876	.2895	.3978	.5132	.6367	.7696	.9133	1.0698	1.2415
1.85	.0180	.0920	.1890	.2917	.4007	.5169	.6413	.7750	.9196	1.0771	1.2497
1.90	.0181	.0927	.1904	.2938	.4036	.5206	.6458	.7804	.9259	1.0842	1.2579
1.95	.0183	.0933	.1918	.2960	.4065	.5243	.6502	.7857	.9321	1.0913	1.2660

TABLE A-5. (CONTINUED)

VALUES OF LAMBDA FOR COHEN'S METHOD

	Percentage of Non-detects										
γ	.01	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
2.00	.0184	.0940	.1932	.2981	.4093	.5279	.6547	.7909	.9382	1.0984	1.2739
2.05	.0186	.0947	.1945	.3001	.4122	.5315	.6590	.7961	.9442	1.1053	1.2819
2.10	.0187	.0954	.1959	.3022	.4149	.5350	.6634	.8013	.9502	1.1122	1.2897
2.15	.0188	.0960	.1972	.3042	.4177	.5385	.6676	.8063	.9562	1.1190	1.2974
2.20	.0189	.0967	.1986	.3062	.4204	.5420	.6719	.8114	.9620	1.1258	1.3051
2.25	.0191	.0973	.1999	.3082	.4231	.5454	.6761	.8164	.9679	1.1325	1.3127
2.30	.0192	.0980	.2012	.3102	.4258	.5488	.6802	.8213	.9736	1.1391	1.3203
2.35	.0193	.0986	.2025	.3122	.4285	.5522	.6844	.8262	.9794	1.1457	1.3278
2.40	.0194	.0992	.2037	.3141	.4311	.5555	.6884	.8311	.9850	1.1522	1.3352
2.45	.0196	.0998	.2050	.3160	.4337	.5588	.6925	.8359	.9906	1.1587	1.3425
2.50	.0197	.1005	.2062	.3179	.4363	.5621	.6965	.8407	.9962	1.1651	1.3498
2.55	.0198	.1011	.2075	.3198	.4388	.5654	.7005	.8454	1.0017	1.1714	1.3571
2.60	.0199	.1017	.2087	.3217	.4414	.5686	.7044	.8501	1.0072	1.1777	1.3642
2.65	.0201	.1023	.2099	.3236	.4439	.5718	.7083	.8548	1.0126	1.1840	1.3714
2.70	.0202	.1029	.2111	.3254	.4464	.5750	.7122	.8594	1.0180	1.1902	1.3784
2.75	.0203	.1035	.2123	.3272	.4489	.5781	.7161	.8639	1.0234	1.1963	1.3854
2.80	.0204	.1040	.2135	.3290	.4513	.5812	.7199	.8685	1.0287	1.2024	1.3924
2.85	.0205	.1046	.2147	.3308	.4537	.5843	.7237	.8730	1.0339	1.2085	1.3993
2.90	.0206	.1052	.2158	.3326	.4562	.5874	.7274	.8775	1.0392	1.2145	1.4061
2.95	.0207	.1058	.2170	.3344	.4585	.5905	.7311	.8819	1.0443	1.2205	1.4129
3.00	.0209	.1063	.2182	.3361	.4609	.5935	.7348	.8863	1.0495	1.2264	1.4197
3.05	.0210	.1069	.2193	.3378	.4633	.5965	.7385	.8907	1.0546	1.2323	1.4264
3.10	.0211	.1074	.2204	.3396	.4656	.5995	.7422	.8950	1.0597	1.2381	1.4330
3.15	.0212	.1080	.2216	.3413	.4679	.6024	.7458	.8993	1.0647	1.2439	1.4396
3.20	.0213	.1085	.2227	.3430	.4703	.6054	.7494	.9036	1.0697	1.2497	1.4462
3.25	.0214	.1091	.2238	.3447	.4725	.6083	.7529	.9079	1.0747	1.2554	1.4527
3.30	.0215	.1096	.2249	.3464	.4748	.6112	.7565	.9121	1.0796	1.2611	1.4592
3.35	.0216	.1102	.2260	.3480	.4771	.6141	.76	.9163	1.0845	1.2668	1.4657
3.40	.0217	.1107	.2270	.3497	.4793	.6169	.7635	.9205	1.0894	1.2724	1.4720
3.45	.0218	.1112	.2281	.3513	.4816	.6197	.7670	.9246	1.0942	1.2779	1.4784
3.50	.0219	.1118	.2292	.3529	.4838	.6226	.7704	.9287	1.0990	1.2835	1.4847
3.55	.0220	.1123	.2303	.3546	.4860	.6254	.7739	.9328	1.1038	1.2890	1.4910
3.60	.0221	.1128	.2313	.3562	.4882	.6282	.7773	.9369	1.1086	1.2945	1.4972
3.65	.0222	.1133	.2324	.3578	.4903	.6309	.7807	.9409	1.1133	1.2999	1.5034
3.70	.0223	.1138	.2334	.3594	.4925	.6337	.7840	.9449	1.1180	1.3053	1.5096
3.75	.0224	.1143	.2344	.3609	.4946	.6364	.7874	.9489	1.1226	1.3107	1.5157
3.80	.0225	.1148	.2355	.3625	.4968	.6391	.7907	.9529	1.1273	1.3160	1.5218
3.85	.0226	.1153	.2365	.3641	.4989	.6418	.7940	.9568	1.1319	1.3213	1.5279
3.90	.0227	.1158	.2375	.3656	.5010	.6445	.7973	.9607	1.1364	1.3266	1.5339
3.95	.0228	.1163	.2385	.3672	.5031	.6472	.8006	.9646	1.1410	1.3318	1.5399

TABLE A-5. (CONTINUED)

VALUES OF LAMBDA FOR COHEN'S METHOD

	Percentage of Non-detects										
γ	.01	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
4.00	.0229	.1168	.2395	.3687	.5052	.6498	.8038	.9685	1.1455	1.3371	1.5458
4.05	.0230	.1173	.2405	.3702	.5072	.6525	.8070	.9723	1.1500	1.3423	1.5518
4.10	.0231	.1178	.2415	.3717	.5093	.6551	.8102	.9762	1.1545	1.3474	1.5577
4.15	.0232	.1183	.2425	.3732	.5113	.6577	.8134	.9800	1.1590	1.3526	1.5635
4.20	.0233	.1188	.2435	.3747	.5134	.6603	.8166	.9837	1.1634	1.3577	1.5693
4.25	.0234	.1193	.2444	.3762	.5154	.6629	.8198	.9875	1.1678	1.3627	1.5751
4.30	.0235	.1197	.2454	.3777	.5174	.6654	.8229	.9913	1.1722	1.3678	1.5809
4.35	.0236	.1202	.2464	.3792	.5194	.6680	.8260	.9950	1.1765	1.3728	1.5866
4.40	.0237	.1207	.2473	.3806	.5214	.6705	.8291	.9987	1.1809	1.3778	1.5924
4.45	.0238	.1212	.2483	.3821	.5234	.6730	.8322	1.0024	1.1852	1.3828	1.5980
4.50	.0239	.1216	.2492	.3836	.5253	.6755	.8353	1.0060	1.1895	1.3878	1.6037
4.55	.0240	.1221	.2502	.3850	.5273	.6780	.8384	1.0097	1.1937	1.3927	1.6093
4.60	.0241	.1225	.2511	.3864	.5292	.6805	.8414	1.0133	1.1980	1.3976	1.6149
4.65	.0241	.1230	.2521	.3879	.5312	.6830	.8445	1.0169	1.2022	1.4024	1.6205
4.70	.0242	.1235	.2530	.3893	.5331	.6855	.8475	1.0205	1.2064	1.4073	1.6260
4.75	.0243	.1239	.2539	.3907	.5350	.6879	.8505	1.0241	1.2106	1.4121	1.6315
4.80	.0244	.1244	.2548	.3921	.5370	.6903	.8535	1.0277	1.2148	1.4169	1.6370
4.85	.0245	.1248	.2558	.3935	.5389	.6928	.8564	1.0312	1.2189	1.4217	1.6425
4.90	.0246	.1253	.2567	.3949	.5407	.6952	.8594	1.0348	1.2230	1.4265	1.6479
4.95	.0247	.1257	.2576	.3963	.5426	.6976	.8623	1.0383	1.2272	1.4312	1.6533
5.00	.0248	.1262	.2585	.3977	.5445	.7000	.8653	1.0418	1.2312	1.4359	1.6587
5.05	.0249	.1266	.2594	.3990	.5464	.7024	.8682	1.0452	1.2353	1.4406	1.6641
5.10	.0249	.1270	.2603	.4004	.5482	.7047	.8711	1.0487	1.2394	1.4453	1.6694
5.15	.0250	.1275	.2612	.4018	.5501	.7071	.8740	1.0521	1.2434	1.4500	1.6747
5.20	.0251	.1279	.2621	.4031	.5519	.7094	.8768	1.0556	1.2474	1.4546	1.6800
5.25	.0252	.1284	.2629	.4045	.5537	.7118	.8797	1.0590	1.2514	1.4592	1.6853
5.30	.0253	.1288	.2638	.4058	.5556	.7141	.8825	1.0624	1.2554	1.4638	1.6905
5.35	.0254	.1292	.2647	.4071	.5574	.7164	.8854	1.0658	1.2594	1.4684	1.6958
5.40	.0255	.1296	.2656	.4085	.5592	.7187	.8882	1.0691	1.2633	1.4729	1.7010
5.45	.0255	.1301	.2664	.4098	.5610	.7210	.8910	1.0725	1.2672	1.4775	1.7061
5.50	.0256	.1305	.2673	.4111	.5628	.7233	.8938	1.0758	1.2711	1.4820	1.7113
5.55	.0257	.1309	.2682	.4124	.5646	.7256	.8966	1.0792	1.2750	1.4865	1.7164
5.60	.0258	.1313	.2690	.4137	.5663	.7278	.8994	1.0825	1.2789	1.4910	1.7215
5.65	.0259	.1318	.2699	.4150	.5681	.7301	.9022	1.0858	1.2828	1.4954	1.7266
5.70	.0260	.1322	.2707	.4163	.5699	.7323	.9049	1.0891	1.2866	1.4999	1.7317
5.75	.0260	.1326	.2716	.4176	.5716	.7346	.9077	1.0924	1.2905	1.5043	1.7368
5.80	.0261	.1330	.2724	.4189	.5734	.7368	.9104	1.0956	1.2943	1.5087	1.7418
5.85	.0262	.1334	.2732	.4202	.5751	.7390	.9131	1.0989	1.2981	1.5131	1.7468
5.90	.0263	.1338	.2741	.4215	.5769	.7412	.9158	1.1021	1.3019	1.5175	1.7518
5.95	.0264	.1342	.2749	.4227	.5786	.7434	.9185	1.1053	1.3057	1.5218	1.7568
6.00	.0264	.1346	.2757	.4240	.5803	.7456	.9212	1.1085	1.3094	1.5262	1.7617

TABLE A-6.

MINIMUM COVERAGE (BETA) OF 95% CONFIDENCE NON-PARAMETRIC UPPER TOLERANCE LIMITS

N	β(maximum)	β(2nd largest)
1	5.0	----
2	22.4	2.6
3	36.8	13.6
4	47.3	24.8
5	54.9	34.2
6	60.7	41.8
7	65.2	48.0
8	68.8	53.0
9	71.7	57.0
10	74.1	60.6
11	76.2	63.6
12	77.9	66.2
13	79.4	68.4
14	80.7	70.4
15	81.9	72.0
16	82.9	73.6
17	83.8	75.0
18	84.7	76.2
19	85.4	77.4
20	86.1	78.4
21	86.7	79.4
22	87.3	80.2
23	87.8	81.0
24	88.3	81.8
25	88.7	82.4
26	89.1	83.0
27	89.5	83.6
28	89.9	84.2
29	90.2	84.6
30	90.5	85.2
31	90.8	85.6
32	91.1	86.0
33	91.3	86.4
34	91.6	86.8
35	91.8	87.2
36	92.0	87.4
37	92.2	87.8
38	92.4	88.2
39	92.6	88.4
40	92.8	88.6

TABLE A-6. (CONTINUED)

MINIMUM COVERAGE (BETA) OF 95% CONFIDENCE NON-PARAMETRIC UPPER TOLERANCE LIMITS

N	β(maximum)	β(2nd largest)
41	93.0	89.0
42	93.1	89.2
43	93.3	89.4
44	93.4	89.6
45	93.6	89.8
46	93.7	90.0
47	93.8	90.2
48	93.9	90.4
49	94.1	90.6
50	94.2	90.8
55	94.7	91.6
60	95.1	92.4
65	95.5	93.0
70	95.8	93.4
75	96.1	93.8
80	96.3	94.2
85	96.5	94.6
90	96.7	94.8
95	96.9	95.0
100	97.0	95.4

TABLE A-7.

CONFIDENCE LEVELS FOR NON-PARAMETRIC PREDICTION LIMITS FOR N=1(1)100

	NUMBER OF FUTURE SAMPLES							
N	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8
1	50.0	33.3	25.0	20.0	16.7	14.3	12.5	11.1
2	66.7	50.0	40.0	33.3	28.6	25.0	22.2	20.0
3	75.0	60.0	50.0	42.9	37.5	33.3	30.0	27.3
4	80.0	66.7	57.1	50.0	44.4	40.0	36.4	33.3
5	83.3	71.4	62.5	55.6	50.0	45.5	41.7	38.5
6	85.7	75.0	66.7	60.0	54.5	50.0	46.2	42.9
7	87.5	77.8	70.0	63.6	58.3	53.8	50.0	46.7

8	88.9	80.0	72.7	66.7	61.5	57.1	53.3	50.0
9	90.0	81.8	75.0	69.2	64.3	60.0	56.3	52.9
10	90.9	83.3	76.9	71.4	66.7	62.5	58.8	55.6
11	91.7	84.6	78.6	73.3	68.8	64.7	61.1	57.9
12	92.3	85.7	80.0	75.0	70.6	66.7	63.2	60.0
13	92.9	86.7	81.3	76.5	72.2	68.4	65.0	61.9
14	93.3	87.5	82.4	77.8	73.7	70.0	66.7	63.6
15	93.8	88.2	83.3	78.9	75.0	71.4	68.2	65.2
16	94.1	88.9	84.2	80.0	76.2	72.7	69.6	66.7
17	94.4	89.5	85.0	81.0	77.3	73.9	70.8	68.0
18	94.7	90.0	85.7	81.8	78.3	75.0	72.0	69.2
19	95.0	90.5	86.4	82.6	79.2	76.0	73.1	70.4
20	95.2	90.9	87.0	83.3	80.0	76.9	74.1	71.4
21	95.5	91.3	87.5	84.0	80.8	77.8	75.0	72.4
22	95.7	91.7	88.0	84.6	81.5	78.6	75.9	73.3
23	95.8	92.0	88.5	85.2	82.1	79.3	76.7	74.2
24	96.0	92.3	88.9	85.7	82.8	80.0	77.4	75.0
25	96.2	92.6	89.3	86.2	83.3	80.6	78.1	75.8
26	96.3	92.9	89.7	86.7	83.9	81.3	78.8	76.5
27	96.4	93.1	90.0	87.1	84.4	81.8	79.4	77.1
28	96.6	93.3	90.3	87.5	84.8	82.4	80.0	77.8
29	96.7	93.5	90.6	87.9	85.3	82.9	80.6	78.4
30	96.8	93.8	90.9	88.2	85.7	83.3	81.1	78.9
31	96.9	93.9	91.2	88.6	86.1	83.8	81.6	79.5
32	97.0	94.1	91.4	88.9	86.5	84.2	82.1	80.0
33	97.1	94.3	91.7	89.2	86.8	84.6	82.5	80.5
34	97.1	94.4	91.9	89.5	87.2	85.0	82.9	81.0
35	97.2	94.6	92.1	89.7	87.5	85.4	83.3	81.4
36	97.3	94.7	92.3	90.0	87.8	85.7	83.7	81.8
37	97.4	94.9	92.5	90.2	88.1	86.0	84.1	82.2
38	97.4	95.0	92.7	90.5	88.4	86.4	84.4	82.6
39	97.5	95.1	92.9	90.7	88.6	86.7	84.8	83.0
40	97.6	95.2	93.0	90.9	88.9	87.0	85.1	83.3

TABLE A-7. (CONTINUED)

CONFIDENCE LEVELS FOR NON-PARAMETRIC PREDICTION LIMITS FOR N=1(1)100

	NUMBER OF FUTURE SAMPLES							
N	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8
41	97.6	95.3	93.2	91.1	89.1	87.2	85.4	83.7
42	97.7	95.5	93.3	91.3	89.4	87.5	85.7	84.0
43	97.7	95.6	93.5	91.5	89.6	87.8	86.0	84.3
44	97.8	95.7	93.6	91.7	89.8	88.0	86.3	84.6
45	97.8	95.7	93.8	91.8	90.0	88.2	86.5	84.9
46	97.9	95.8	93.9	92.0	90.2	88.5	86.8	85.2
47	97.9	95.9	94.0	92.2	90.4	88.7	87.0	85.5
48	98.0	96.0	94.1	92.3	90.6	88.9	87.3	85.7
49	98.0	96.1	94.2	92.5	90.7	89.1	87.5	86.0
50	98.0	96.2	94.3	92.6	90.9	89.3	87.7	86.2
51	98.1	96.2	94.4	92.7	91.1	89.5	87.9	86.4
52	98.1	96.3	94.5	92.9	91.2	89.7	88.1	86.7
53	98.1	96.4	94.6	93.0	91.4	89.8	88.3	86.9
54	98.2	96.4	94.7	93.1	91.5	90.0	88.5	87.1
55	98.2	96.5	94.8	93.2	91.7	90.2	88.7	87.3
56	98.2	96.6	94.9	93.3	91.8	90.3	88.9	87.5
57	98.3	96.6	95.0	93.4	91.9	90.5	89.1	87.7
58	98.3	96.7	95.1	93.5	92.1	90.6	89.2	87.9
59	98.3	96.7	95.2	93.7	92.2	90.8	89.4	88.1
60	98.4	96.8	95.2	93.8	92.3	90.9	89.6	88.2
61	98.4	96.8	95.3	93.8	92.4	91.0	89.7	88.4
62	98.4	96.9	95.4	93.9	92.5	91.2	89.9	88.6
63	98.4	96.9	95.5	94.0	92.6	91.3	90.0	88.7
64	98.5	97.0	95.5	94.1	92.8	91.4	90.1	88.9
65	98.5	97.0	95.6	94.2	92.9	91.5	90.3	89.0
66	98.5	97.1	95.7	94.3	93.0	91.7	90.4	89.2
67	98.5	97.1	95.7	94.4	93.1	91.8	90.5	89.3
68	98.6	97.1	95.8	94.4	93.2	91.9	90.7	89.5
69	98.6	97.2	95.8	94.5	93.2	92.0	90.8	89.6
70	98.6	97.2	95.9	94.6	93.3	92.1	90.9	89.7
71	98.6	97.3	95.9	94.7	93.4	92.2	91.0	89.9
72	98.6	97.3	96.0	94.7	93.5	92.3	91.1	90.0
73	98.6	97.3	96.1	94.8	93.6	92.4	91.3	90.1
74	98.7	97.4	96.1	94.9	93.7	92.5	91.4	90.2
75	98.7	97.4	96.2	94.9	93.8	92.6	91.5	90.4
76	98.7	97.4	96.2	95.0	93.8	92.7	91.6	90.5
77	98.7	97.5	96.3	95.1	93.9	92.8	91.7	90.6

78	98.7	97.5	96.3	95.1	94.0	92.9	91.8	90.7
79	98.8	97.5	96.3	95.2	94.0	92.9	91.9	90.8
80	98.8	97.6	96.4	95.2	94.1	93.0	92.0	90.9

TABLE A-7. (CONTINUED)

CONFIDENCE LEVELS FOR NON-PARAMETRIC PREDICTION LIMITS FOR N=1(1)100

	NUMBER OF FUTURE SAMPLES							
N	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8
81	98.8	97.6	96.4	95.3	94.2	93.1	92.0	91.0
82	98.8	97.6	96.5	95.3	94.3	93.2	92.1	91.1
83	98.8	97.6	96.5	95.4	94.3	93.3	92.2	91.2
84	98.8	97.7	96.6	95.5	94.4	93.3	92.3	91.3
85	98.8	97.7	96.6	95.5	94.4	93.4	92.4	91.4
86	98.9	97.7	96.6	95.6	94.5	93.5	92.5	91.5
87	98.9	97.8	96.7	95.6	94.6	93.5	92.6	91.6
88	98.9	97.8	96.7	95.7	94.6	93.6	92.6	91.7
89	98.9	97.8	96.7	95.7	94.7	93.7	92.7	91.8
90	98.9	97.8	96.8	95.7	94.7	93.8	92.8	91.8
91	98.9	97.8	96.8	95.8	94.8	93.8	92.9	91.9
92	98.9	97.9	96.8	95.8	94.8	93.9	92.9	92.0
93	98.9	97.9	96.9	95.9	94.9	93.9	93.0	92.1
94	98.9	97.9	96.9	95.9	94.9	94.0	93.1	92.2
95	99.0	97.9	96.9	96.0	95.0	94.1	93.1	92.2
96	99.0	98.0	97.0	96.0	95.0	94.1	93.2	92.3
97	99.0	98.0	97.0	96.0	95.1	94.2	93.3	92.4
98	99.0	98.0	97.0	96.1	95.1	94.2	93.3	92.5
99	99.0	98.0	97.1	96.1	95.2	94.3	93.4	92.5
100	99.0	98.0	97.1	96.2	95.2	94.3	93.5	92.6

I. CONSTRUCTION OF POWER CURVES

To construct power curves for each of the parametric and non-parametric retesting strategies, random standard Normal deviates were generated on an IBM mainframe computer using SAS. The background level mean concentration was set to zero, while the alternative mean concentration level was incremented in steps of Δ=0.5 standardized units above the background level. At each increment, 5000 iterations of the retesting strategy were simulated; the proportion of iterations indicating contamination at any one of the wells in the downgradient monitoring network was designated as the effective power of the retesting strategy (for that Δ and configuration of background samples and monitoring wells).

Power values for the EPA Reference Power Curves were not simulated, but represent analytical calculations based on the non-central t-distribution with non-centrality parameter Δ. SAS programs for simulating the effective power of any of the parametric or non-parametric retesting strategies are presented below.

```
//*******************************************************************;
//*   DESCRIPTION:  *** PARAMETRIC SIMULATIONS ***
//*
//*   This program produces power curves for 35 different curve
//*   simulations (refer to the %LET statements below).  Delta ranges
//*   from 0 to 5 by 0.5.  The variable list is as follows for the
//*   input parameters:
//*
//*   BG = Background
//*   WL = Well
//*   TL = Tolerance Limit
//*   PL = Prediction Limit
//*
//*******************************************************************;
//    EXEC SAS
//    OUTSAS DD DSN=XXXXXXX.GWT03000.SJA3092.CURVES,
//    DISP=OLD
//    SYSIN DD *

OPTIONS LS=132 PS=57;
%LET ISTART=1;
%LET CURVENUM=35;
%LET RSEED=2020;
%LET REPEAT=5000;
%LET ITPRINT=1000;

%LET BG1 =24;      %LET WL1 =5;       %LET TL1 =0.95;    %LET PL1 =0.80;
%LET BG2 =24;      %LET WL2 =5;       %LET TL2 =0.95;    %LET PL2 =0.85;
%LET BG3 =8;       %LET WL3 =5;       %LET TL3 =0.95;    %LET PL3 =0.80;
%LET BG4 =8;       %LET WL4 =5;       %LET TL4 =0.95;    %LET PL4 =0.85;
%LET BG5 =24;      %LET WL5 =20;      %LET TL5 =0.95;    %LET PL5 =0.95;
%LET BG6 =24;      %LET WL6 =20;      %LET TL6 =0.95;    %LET PL6 =0.97;
%LET BG7 =8;       %LET WL7 =20;      %LET TL7 =0.95;    %LET PL7 =0.95;
%LET BG8 =8;       %LET WL8 =20;      %LET TL8 =0.95;    %LET PL8 =0.97;
%LET BG9 =24;      %LET WL9 =50;      %LET TL9 =0.95;    %LET PL9 =0.98;
%LET BG10=24;      %LET WL10=50;      %LET TL10=0.95;    %LET PL10=0.99;
```

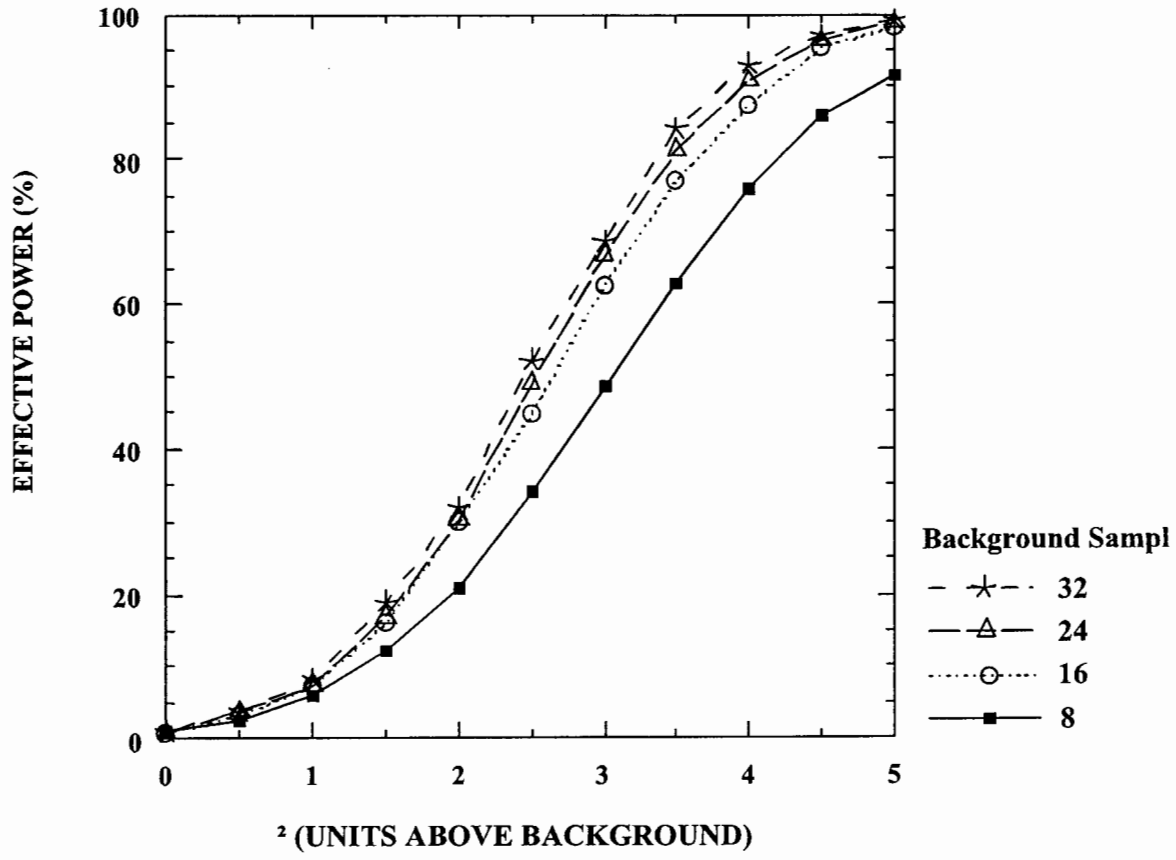
```
%LET BG11=24;      %LET WL11=50;      %LET TL11=0.99;    %LET PL11=0.90;
%LET BG12=24;      %LET WL12=50;      %LET TL12=0.99;    %LET PL12=0.93;
%LET BG13=24;      %LET WL13=50;      %LET TL13=0.99;    %LET PL13=0.94;
%LET BG14=24;      %LET WL14=50;      %LET TL14=0.98;    %LET PL14=0.95;
%LET BG15=24;      %LET WL15=50;      %LET TL15=0.98;    %LET PL15=0.97;
%LET BG16=24;      %LET WL16=100;     %LET TL16=0.98;    %LET PL16=0.97;
%LET BG17=24;      %LET WL17=100;     %LET TL17=0.98;    %LET PL17=0.99;
%LET BG18=24;      %LET WL18=100;     %LET TL18=0.99;    %LET PL18=0.95;
%LET BG19=24;      %LET WL19=100;     %LET TL19=0.99;    %LET PL19=0.97;
%LET BG20=24;      %LET WL20=100;     %LET TL20=0.99;    %LET PL20=0.98;
%LET BG21=8;       %LET WL21=20;      %LET TL21=0.95;    %LET PL21=0.98;
%LET BG22=8;       %LET WL22=5;       %LET TL22=0.95;    %LET PL22=0.90;
%LET BG23=16;      %LET WL23=5;       %LET TL23=0.95;    %LET PL23=0.85;
%LET BG24=16;      %LET WL24=5;       %LET TL24=0.95;    %LET PL24=0.90;
%LET BG25=24;      %LET WL25=5;       %LET TL25=0.95;    %LET PL25=0.90;
%LET BG26=16;      %LET WL26=20;      %LET TL26=0.95;    %LET PL26=0.95;
%LET BG27=16;      %LET WL27=20;      %LET TL27=0.95;    %LET PL27=0.97;
%LET BG28=16;      %LET WL28=50;      %LET TL28=0.98;    %LET PL28=0.95;
%LET BG29=16;      %LET WL29=50;      %LET TL29=0.98;    %LET PL29=0.97;
%LET BG30=16;      %LET WL30=50;      %LET TL30=0.99;    %LET PL30=0.90;
%LET BG31=16;      %LET WL31=50;      %LET TL31=0.99;    %LET PL31=0.92;
%LET BG32=24;      %LET WL32=100;     %LET TL32=0.98;    %LET PL32=0.98;
%LET BG33=16;      %LET WL33=100;     %LET TL33=0.98;    %LET PL33=0.98;
%LET BG34=16;      %LET WL34=100;     %LET TL34=0.99;    %LET PL34=0.95;
%LET BG35=16;      %LET WL35=100;     %LET TL35=0.99;    %LET PL35=0.96;

%MACRO PARSIM;
DATA ITERATE;
*** Set changing simulation variable to common variable names;
      BG=&&BG&I;
      WL=&&WL&I;
      TL=&&TL&I;
      PL=&&PL&I;

DO DELTA=0 TO 5 BY 0.5;
*** Initialize TP0, TP1 & TP2 to 0 before entering simulation;
      TP0=0;
      TP1=0;
      TP2=0;

DO J=1 TO &REPEAT;
*** Initialize CNT0, CNT1 & CNT2 to 0;
      CNT0=0;
      CNT1=0;
      CNT2=0;

XB=RANNOR(&RSEED)/SQRT(BG);
SB=SQRT(2*RANGAM(&RSEED,(BG-1)/2)/(BG-1));

PL2=XB+SB*SQRT(1+1/BG)*TINV((1-(1-PL)/2),(BG-1));
PL1=XB+SB*SQRT(1+1/BG)*TINV((1-(1-PL)),(BG-1));
PL0=XB+SB*SQRT(1+1/BG)*TINV((1-(1-TL)),(BG-1));
TLIM=XB+SB*SQRT(1+1/BG)*TINV((1-(1-TL)),(BG-1));

DO K=1 TO WL;
      IF K<WL THEN DO;
      X1=RANNOR(&RSEED);
      X2=RANNOR(&RSEED);
      X3=RANNOR(&RSEED);
      END;
      ELSE DO;
      X1=RANNOR(&RSEED)+DELTA;
      X2=RANNOR(&RSEED)+DELTA;
      X3=RANNOR(&RSEED)+DELTA;
```

```
      END;
      IF X1>TLIM THEN DO;
      CNT0=CNT0+1;
      IF X2>PL1 THEN CNT1=CNT1+1;
      IF X2>PL2 OR X3>PL2 THEN CNT2=CNT2+1;
      END;
END;

IF CNT0>0 THEN TP0=TP0+100/&REPEAT;
IF CNT1>0 THEN TP1=TP1+100/&REPEAT;
IF CNT2>0 THEN TP2=TP2+100/&REPEAT;

*** Print iteration information every 100 iterations;
I=&I;
IF MOD(J,&ITPRINT)=0 THEN
   PUT '>>> CURVE ' I ', ITERATION ' J ', ' BG= ', ' WL= ', ' TL= ', '
      PL= ', ' DELTA= ', ' TP0= ', ' TP1= ', ' TP2= '<<<';
END;
OUTPUT;
END;
RUN;

DATA OUTSAS.PCURVE&I; SET ITERATE(KEEP=BG WL TL PL TP0 TP1 TP2 DELTA);
RUN;

PROC PRINT DATA=OUTSAS.PCURVE&I;
 FORMAT TP0 TP1 TP2 8.4;
 TITLE1"TEST PRINT OF PARAMETRIC SIMULATION PCURVE&I";
 TITLE2"NUMBER OF ITERATIONS = &REPEAT";
RUN;

%MEND PARSIM;
 %MACRO CURVE;
  %DO I=&ISTART %TO &CURVENUM;
   %PARSIM
  %END;
 %MEND CURVE;
%CURVE


//*******************************************************************;
//*   DESCRIPTION:  *** NON-PARAMETRIC SIMULATION ***
//*
//*   This program produces power curves for 15 different curve
//*   simulations (refer to the %LET statements below).  Delta ranges
//*   from 0 to 5 by 0.5.  The variable list is as follows for the
//*   input parameters:
//*
//*   BG = Background
//*   WL = Well
//*
//*******************************************************************;
//    EXEC SAS
//    OUTSAS DD DSN=XXXXXXX.GWT03000.SJA3092.CURVES,DISP=OLD
//    SYSIN DD *

OPTIONS LS=132 PS=57;
%LET ISTART=1;
%LET CURVENUM=15;
%LET RSEED=3030;
%LET REPEAT=5000;
%LET ITPRINT=1000;

%LET BG1 =8;      %LET WL1 =5;
```

```
%LET BG2 =16;      %LET WL2 =5;
%LET BG3 =24;      %LET WL3 =5;
%LET BG4 =8;       %LET WL4 =20;
%LET BG5 =16;      %LET WL5 =20;
%LET BG6 =24;      %LET WL6 =20;
%LET BG7 =8;       %LET WL7 =50;
%LET BG8 =16;      %LET WL8 =50;
%LET BG9 =24;      %LET WL9 =50;
%LET BG10=8;       %LET WL10=100;
%LET BG11=16;      %LET WL11=100;
%LET BG12=24;      %LET WL12=100;
%LET BG13=32;      %LET WL13=100;
%LET BG14=32;      %LET WL14=20;
%LET BG15=32;      %LET WL15=50;

%MACRO NPARSIM;
DATA ITERATE;
 *** Set changing simulation variable to common variable names;
 BG=&&BG&I;
 WL=&&WL&I;

 DO DELTA=0 TO 5 BY 0.5;
      *** Initialize PLx variables to 0 before entering simulation;
      PL0=0;
      PL1=0;
      PL2A=0;
      PL2B=0;
      PL3A=0;
      PL3B=0;

 DO J=1 TO &REPEAT;
      *** Initialize CNTx variables to 0;
      CNT0=0;
      CNT1=0;
      CNT2=0;
      CNT3=0;
      CNT4=0;
      CNT5=0;

 DO K=1 TO BG;
      TEST=RANNOR(&RSEED);
      IF K=1 THEN MAX=TEST;
       ELSE IF TEST>MAX THEN MAX=TEST;
 END;

 DO L=1 TO WL;
      IF L<WL THEN DO;
      X1=RANNOR(&RSEED);
      X2=RANNOR(&RSEED);
      X3=RANNOR(&RSEED);
      X4=RANNOR(&RSEED);
      END;
      ELSE DO;
      X1=RANNOR(&RSEED)+DELTA;
      X2=RANNOR(&RSEED)+DELTA;
      X3=RANNOR(&RSEED)+DELTA;
      X4=RANNOR(&RSEED)+DELTA;
END;
IF X1>MAX THEN DO;
      CNT0=CNT0+1;
      IF X2>MAX THEN CNT1=CNT1+1;
      IF X2>MAX & X3>MAX THEN CNT2=CNT2+1;
      IF X2>MAX OR X3>MAX THEN CNT3=CNT3+1;
      IF X2>MAX & X3>MAX & X4>MAX THEN CNT4=CNT4+1;
```

```
      IF X2>MAX OR X3>MAX OR X4>MAX THEN CNT5=CNT5+1;
END;

IF CNT0>0 THEN PL0=PL0+100/&REPEAT;
IF CNT1>0 THEN PL1=PL1+100/&REPEAT;
IF CNT2>0 THEN PL2A=PL2A+100/&REPEAT;
IF CNT3>0 THEN PL2B=PL2B+100/&REPEAT;
IF CNT4>0 THEN PL3A=PL3A+100/&REPEAT;
IF CNT5>0 THEN PL3B=PL3B+100/&REPEAT;

*** Print iteration information every X iterations;
I=&I;
IF MOD(J,&ITPRINT)=0 THEN
 PUT '>>> CURVE ' I ', ITERATION ' J ', ' BG= ', ' WL= ', ' DELTA=
      ', ' PL0= ', ' PL1= ', ' PL2A= ', ' PL2B= ', ' PL3A= ', ' PL3B= '<<<';
END;
OUTPUT;
END;
RUN;

DATA OUTSAS.NCURVE&I; SET ITERATE(KEEP=BG WL PL0 PL1 PL2A PL2B PL3A PL3B DELTA);
RUN;

PROC PRINT DATA=OUTSAS.NCURVE&I;
 FORMAT PL0 PL1 PL2A PL2B PL3A PL3B 8.4;
 TITLE1"TEST PRINT OF NON-PARAMETRIC SIMULATION NCURVE&I";
 TITLE2"NUMBER OF ITERATIONS = &REPEAT";
RUN;

%MEND NPARSIM;
 %MACRO CURVE;
  %DO I=&ISTART %TO &CURVENUM;
   %NPARSIM
  %END;
 %MEND CURVE;
%CURVE
```

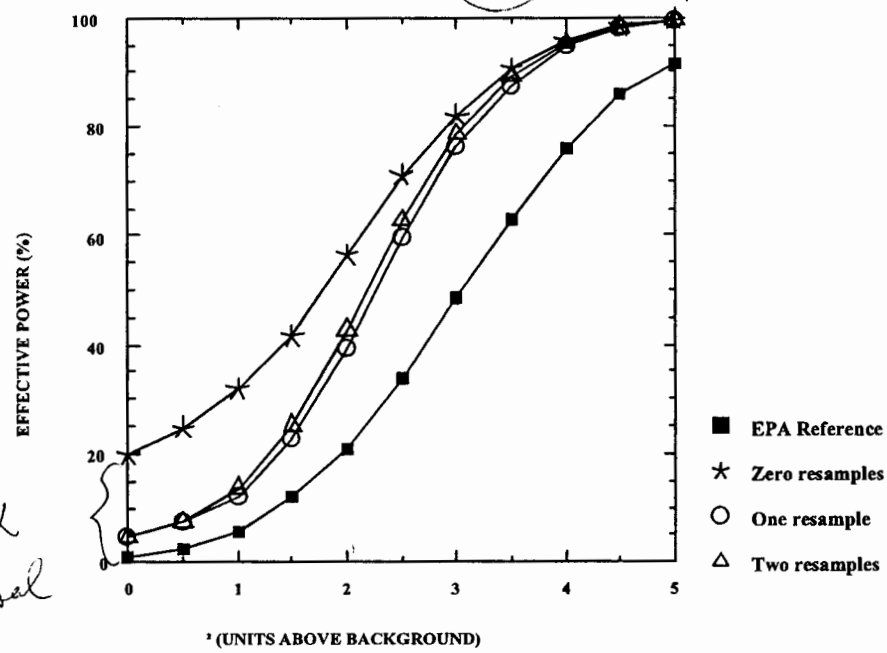
EPA REFERENCE POWER CURVES



II. PARAMETRIC RETESTING STRATEGIES

POWER CURVE FOR 95% TOLERANCE AND 90% PREDICTION LIMIT

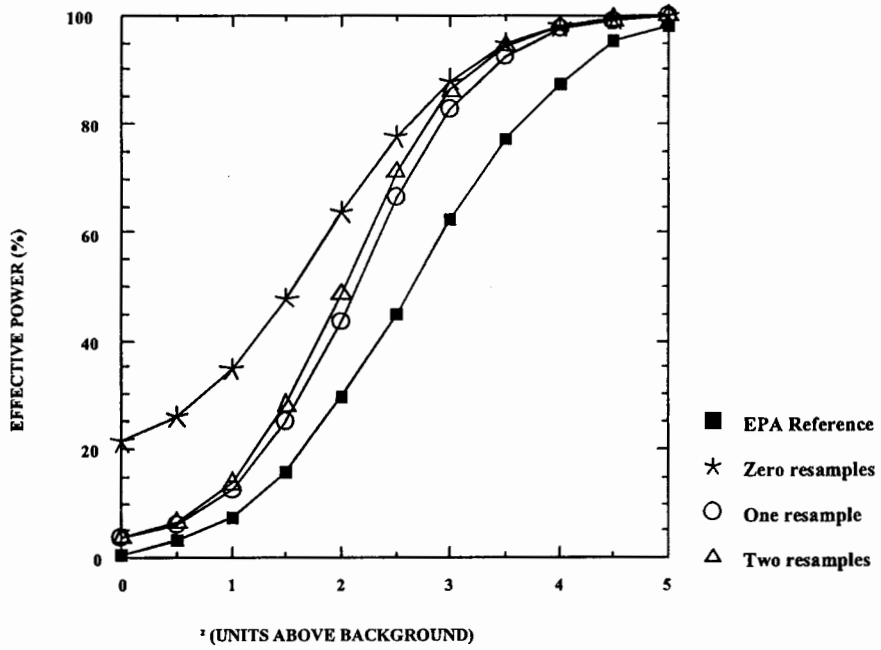
(8 Background Samples, 5 wells) → produce 20% false positive



α total for 5 wells

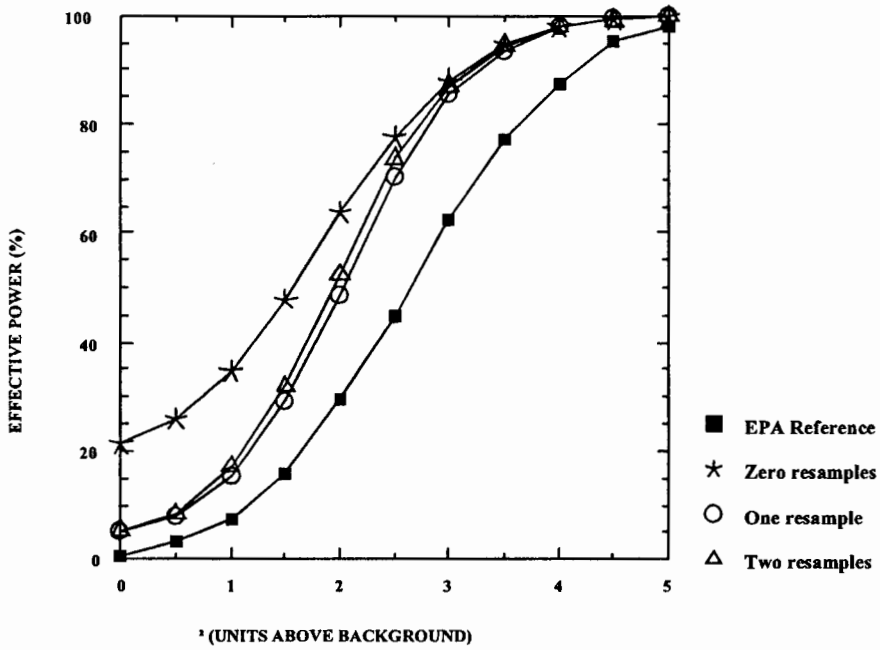
POWER CURVE FOR 95% TOLERANCE AND 90% PREDICTION LIMIT

(16 Background Samples; 5 wells)



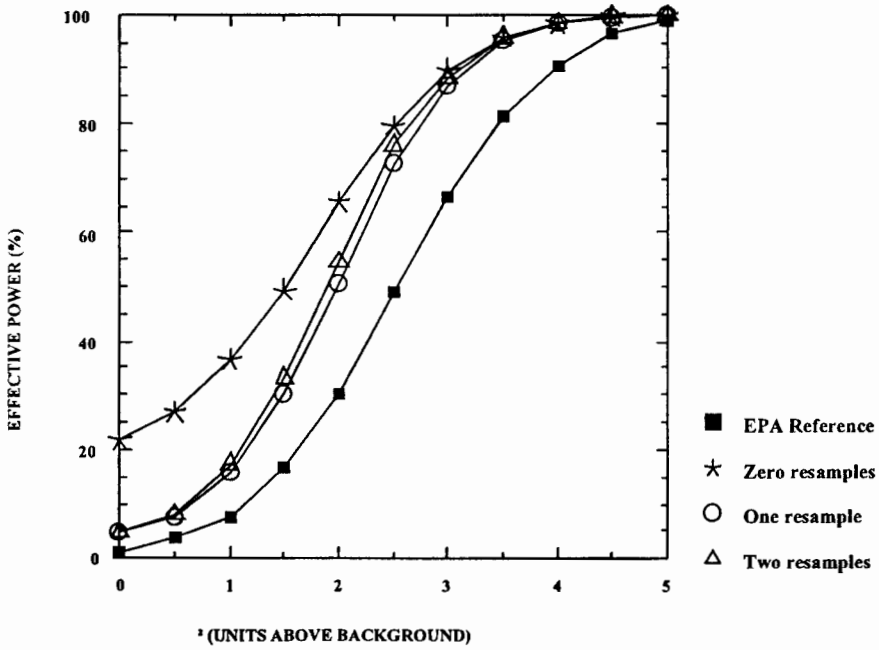
POWER CURVE FOR 95% TOLERANCE AND 85% PREDICTION LIMIT

(16 Background Samples; 5 wells)



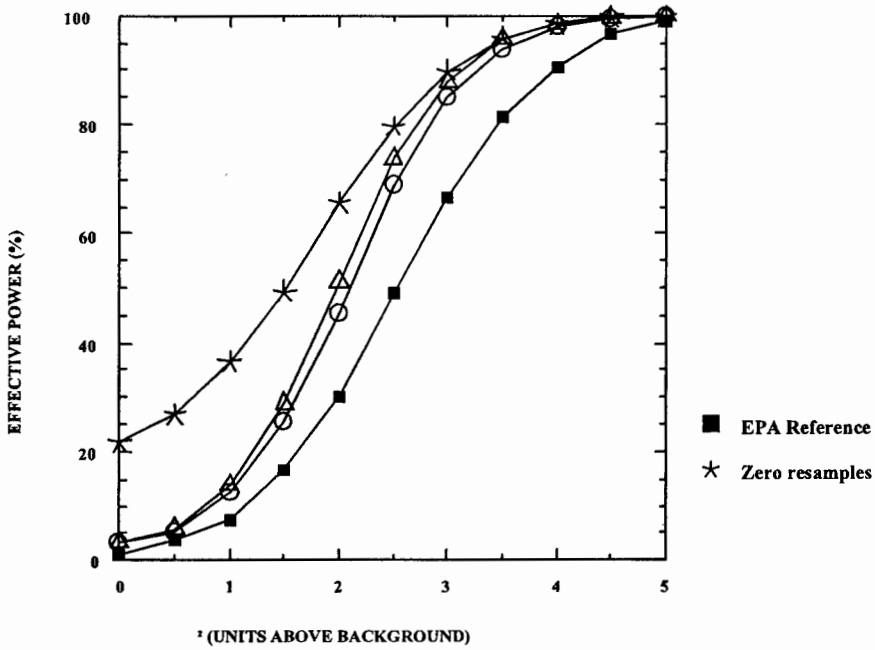
POWER CURVE FOR 95% TOLERANCE AND 85% PREDICTION LIMIT

(24 Background Samples; 5 wells)



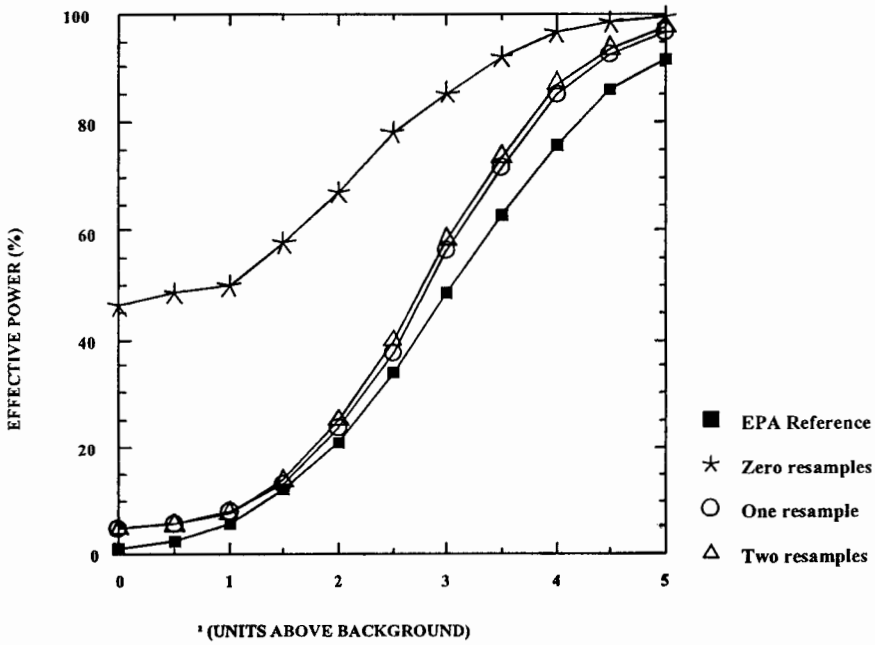
POWER CURVE FOR 95% TOLERANCE AND 90% PREDICTION LIMIT

(24 Background Samples; 5 wells)



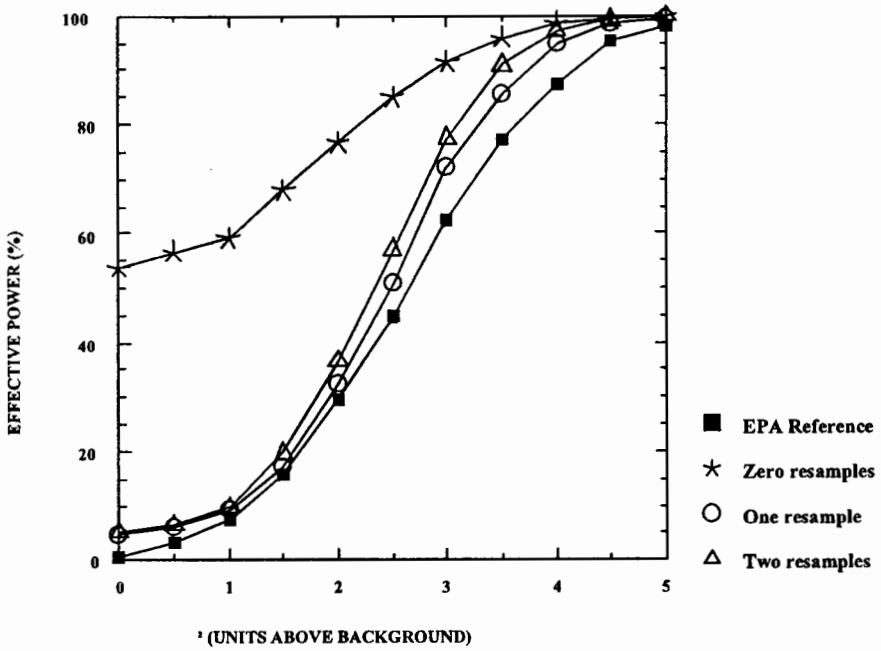
POWER CURVE FOR 95% TOLERANCE AND 98% PREDICTION LIMIT

(8 Background Samples; 20 wells)



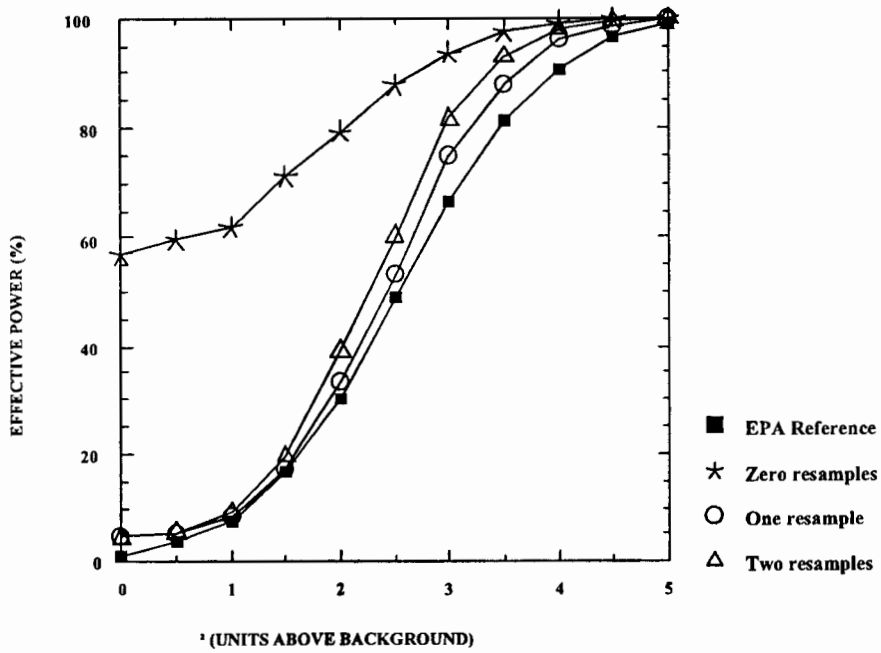
POWER CURVE FOR 95% TOLERANCE AND 97% PREDICTION LIMIT

(16 Background Samples; 20 wells)



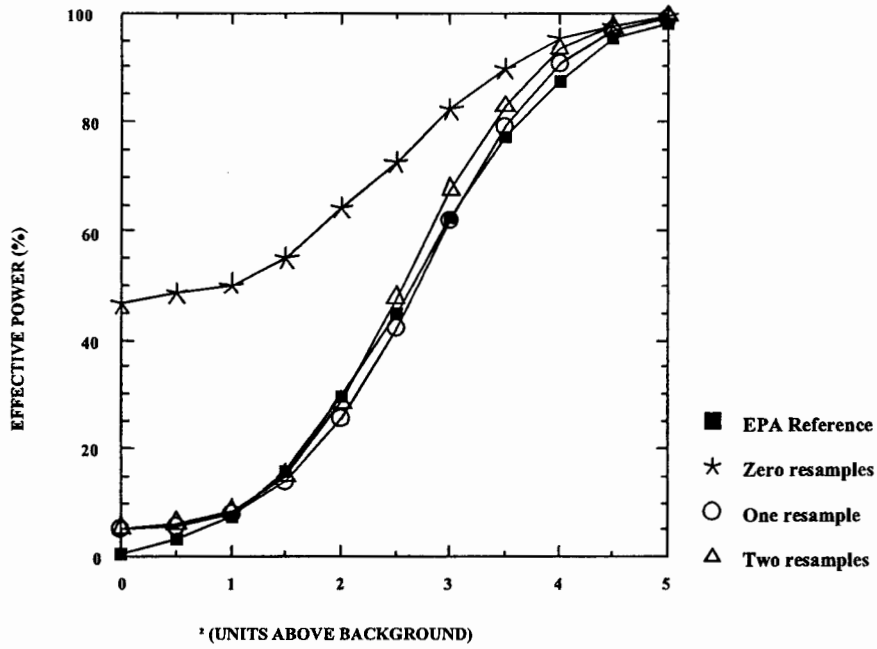
POWER CURVE FOR 95% TOLERANCE AND 97% PREDICTION LIMIT

(24 Background Samples; 20 wells)



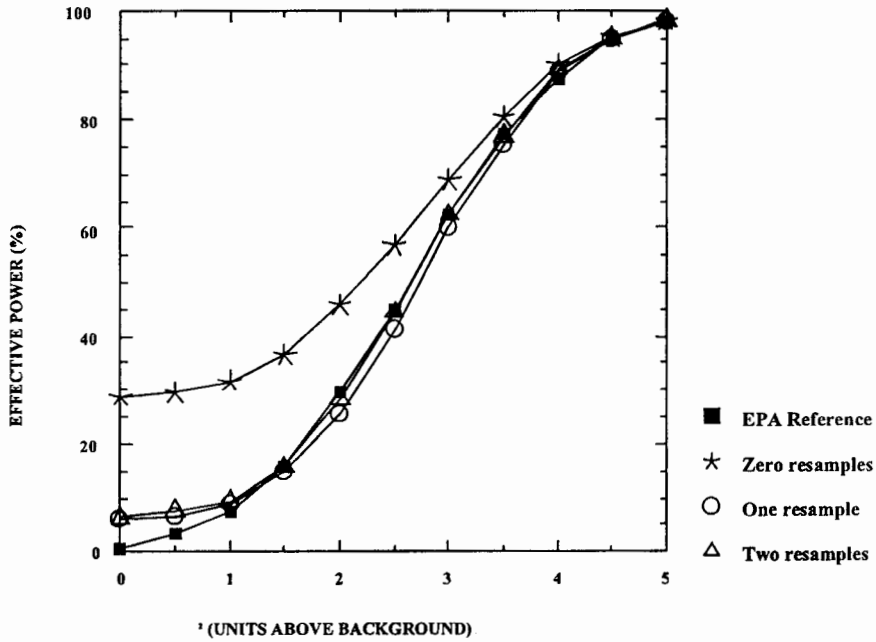
POWER CURVE FOR 98% TOLERANCE AND 97% PREDICTION LIMIT

(16 Background Samples; 50 wells)



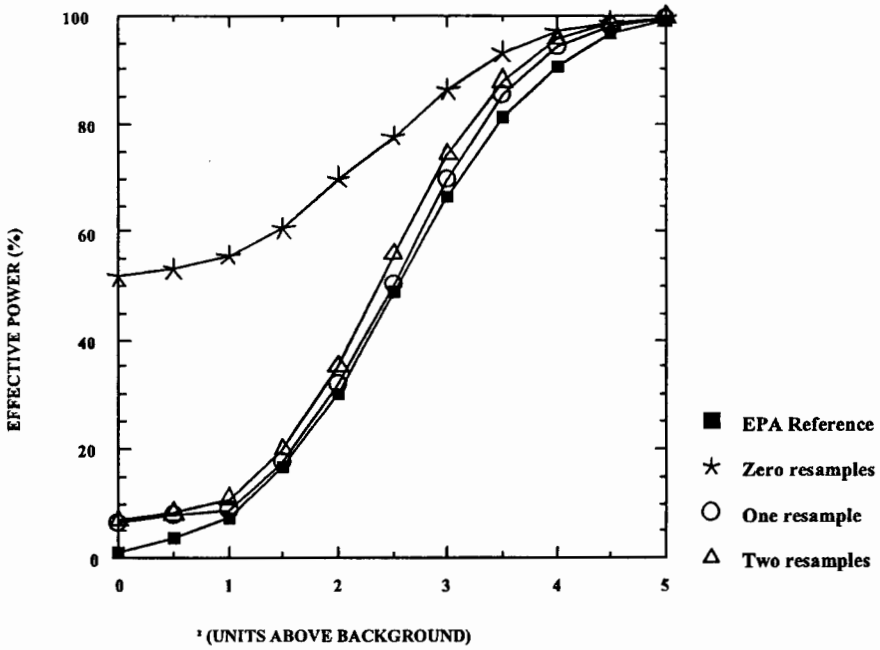
POWER CURVE FOR 99% TOLERANCE AND 92% PREDICTION LIMIT

(16 Background Samples; 50 wells)



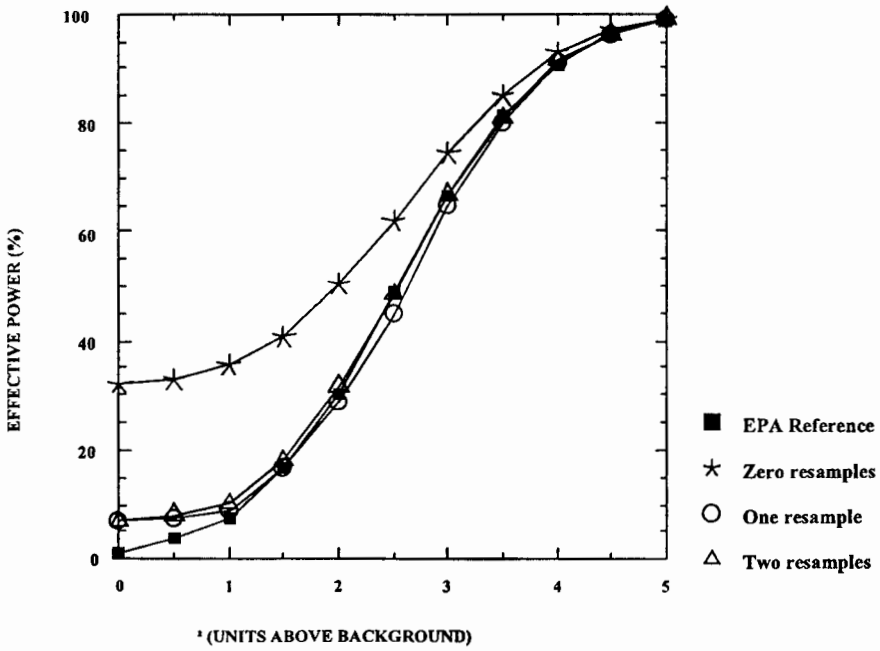
POWER CURVE FOR 98% TOLERANCE AND 95% PREDICTION LIMIT

(24 Background Samples; 50 wells)



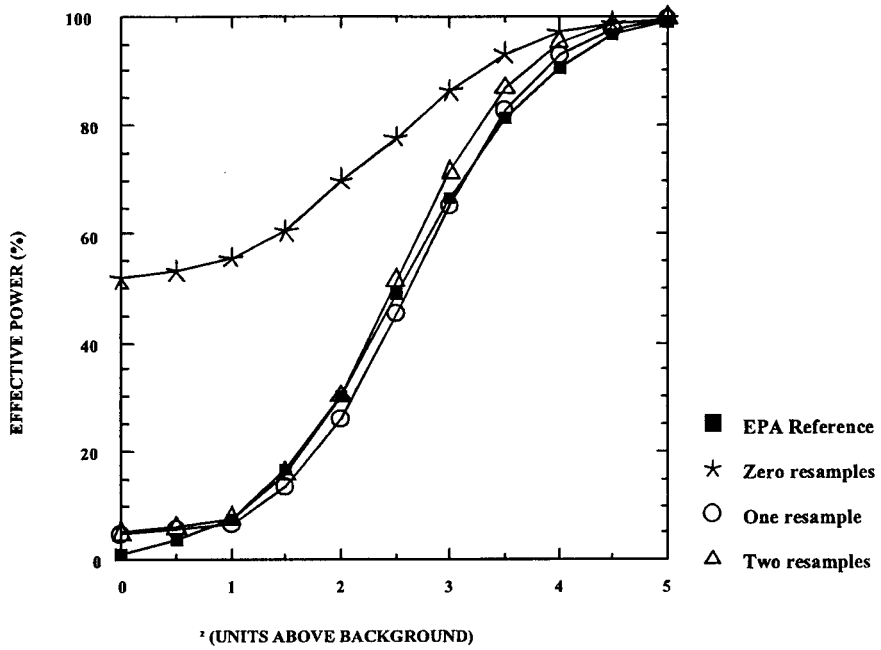
POWER CURVE FOR 99% TOLERANCE AND 90% PREDICTION LIMIT

(24 Background Samples; 50 wells)



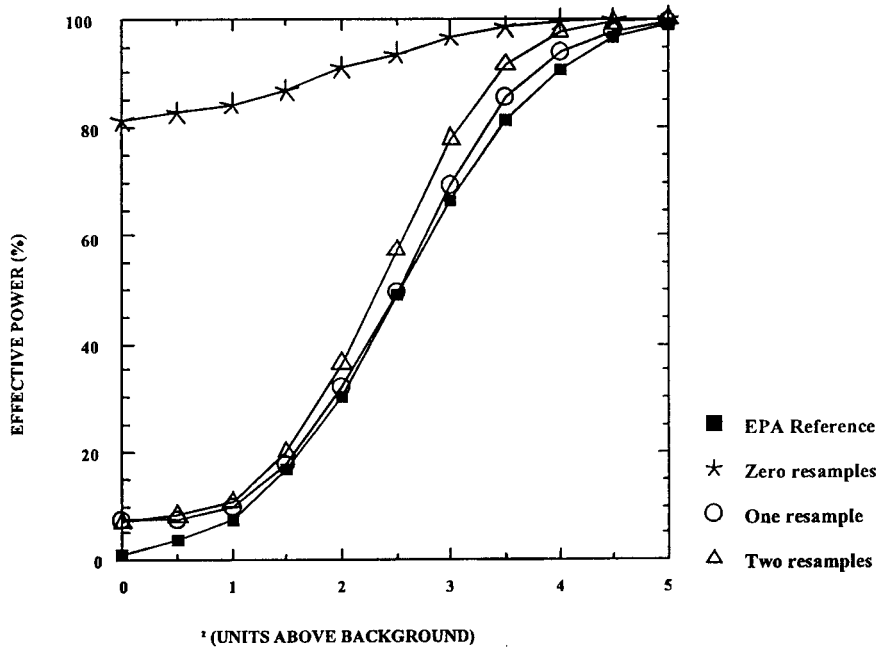
POWER CURVE FOR 98% TOLERANCE AND 97% PREDICTION LIMIT

(24 Background Samples; 50 wells)



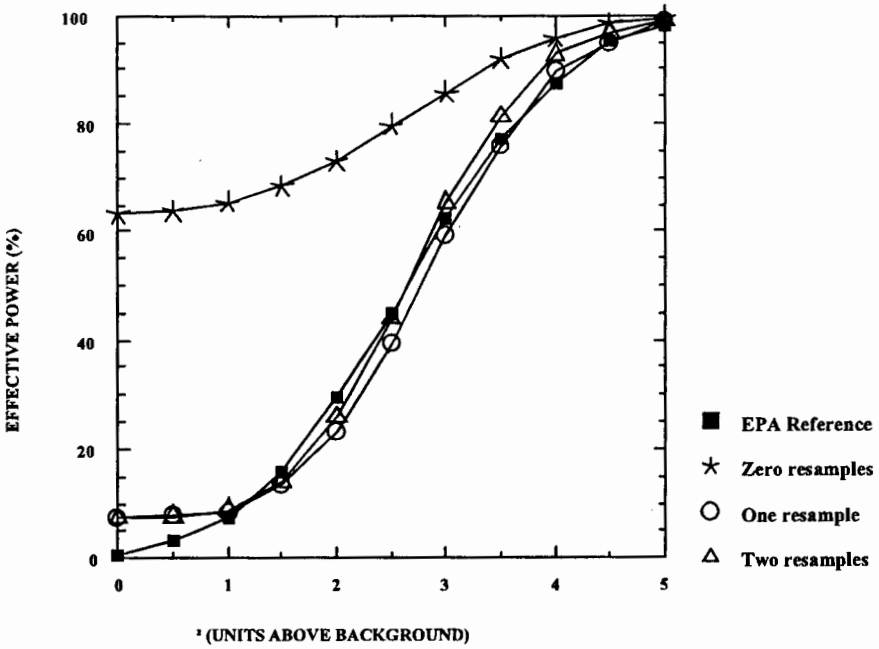
POWER CURVE FOR 95% TOLERANCE AND 98% PREDICTION LIMIT

(24 Background Samples; 50 wells)



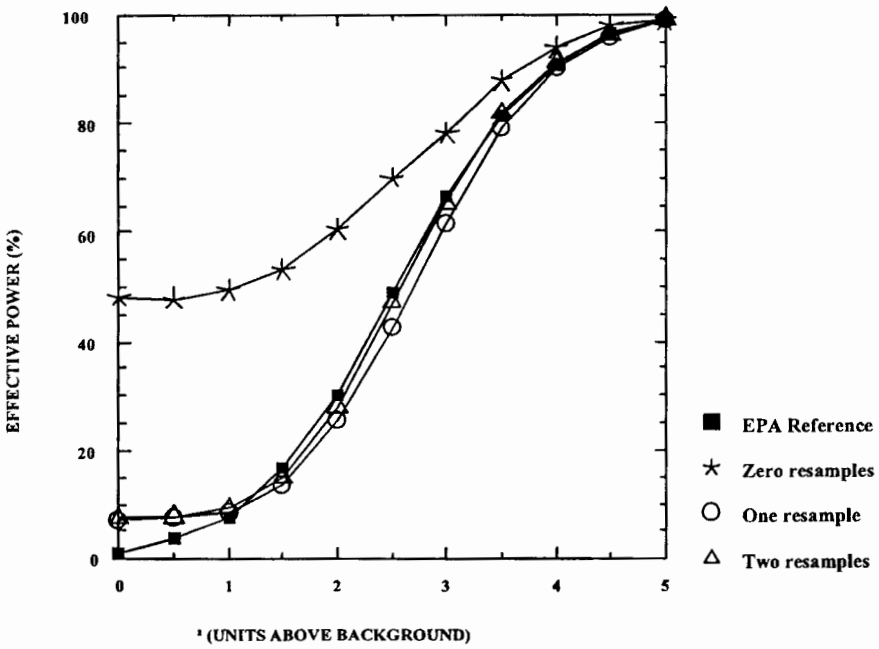
POWER CURVE FOR 98% TOLERANCE AND 98% PREDICTION LIMIT

(16 Background Samples; 100 wells)



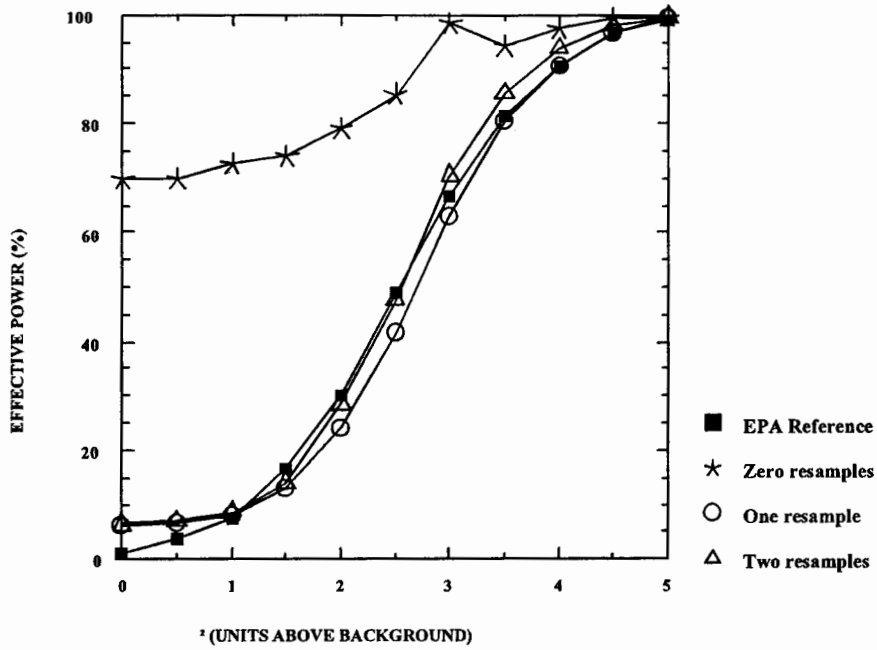
POWER CURVE FOR 99% TOLERANCE AND 95% PREDICTION LIMIT

(24 Background Samples; 100 wells)



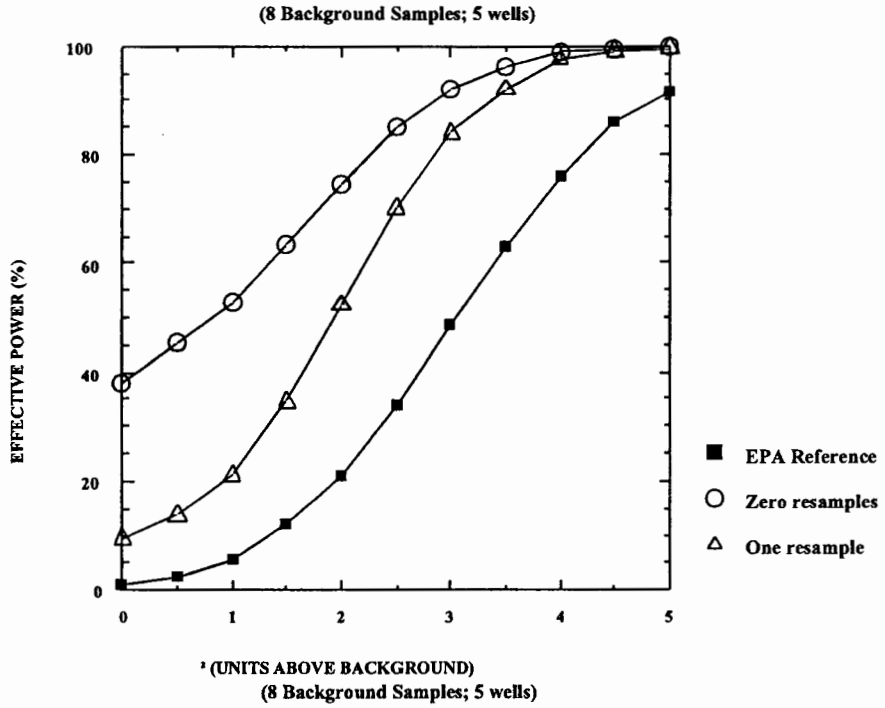
POWER CURVE FOR 98% TOLERANCE AND 98% PREDICTION LIMIT

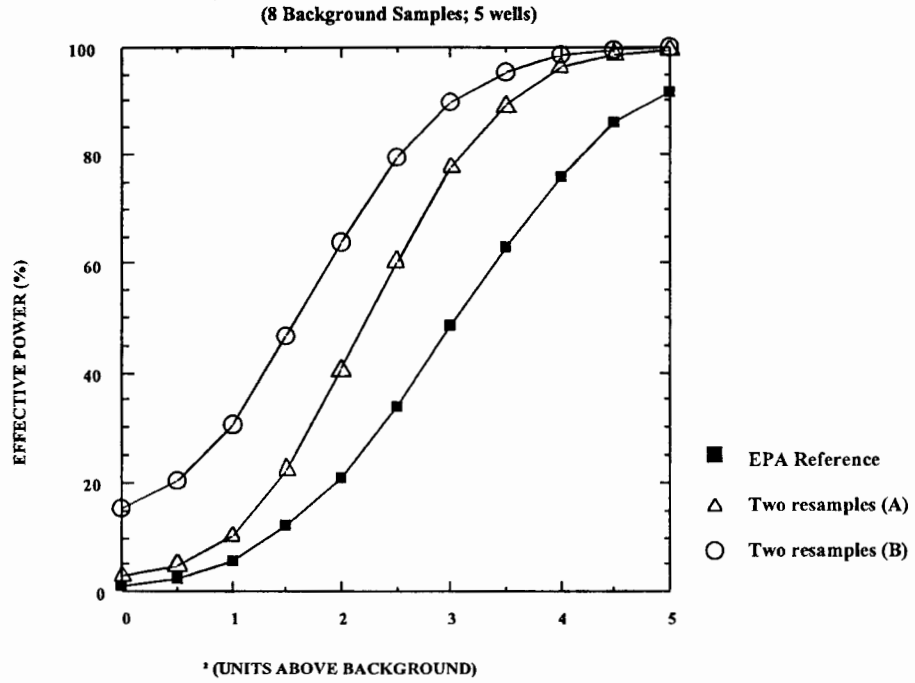
(24 Background Samples; 100 wells)

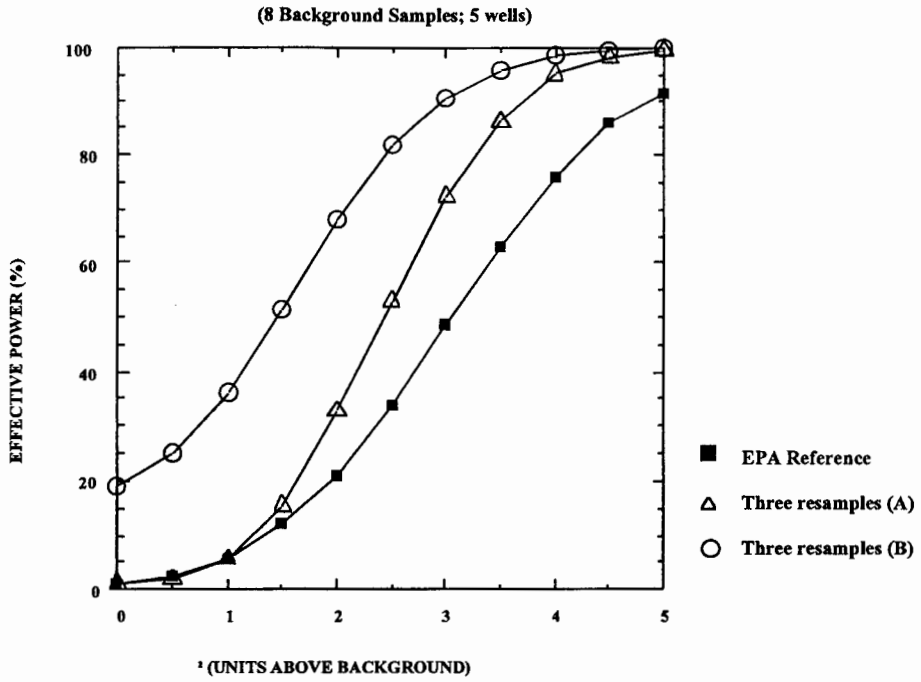


III. NON-PARAMETRIC RETESTING STRATEGIES

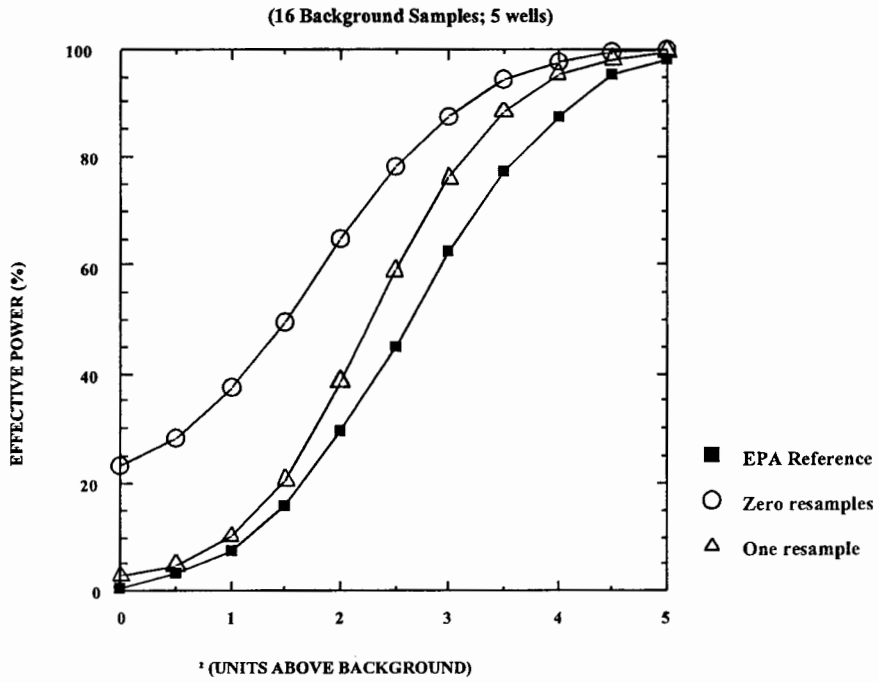
POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

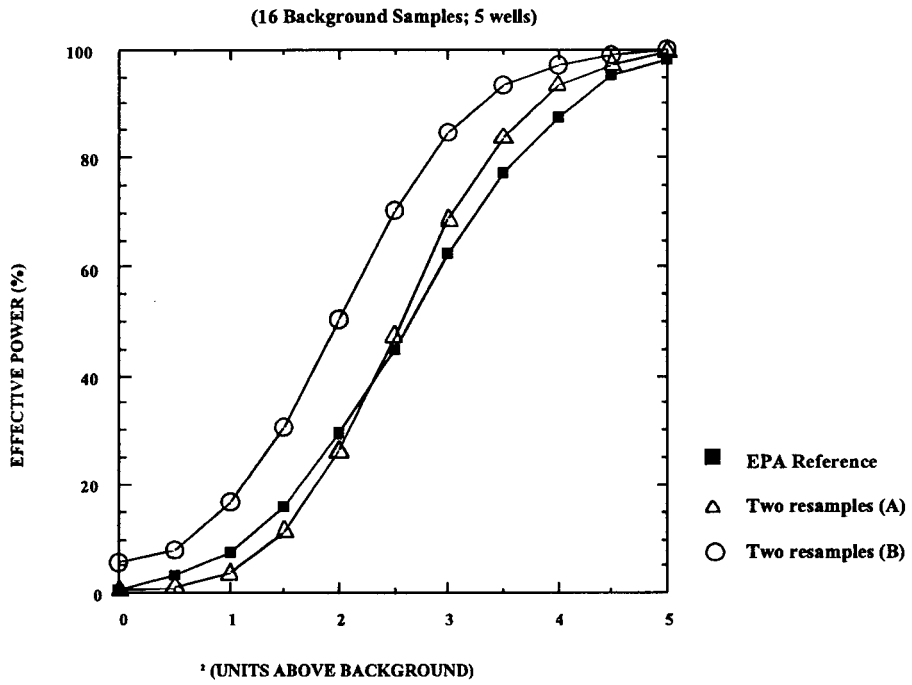


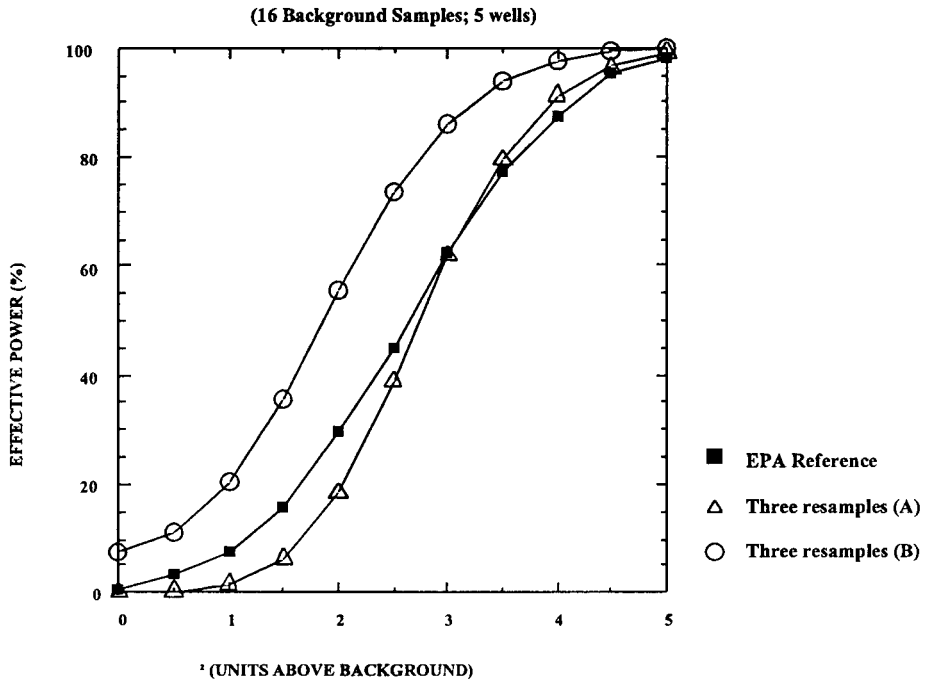




POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

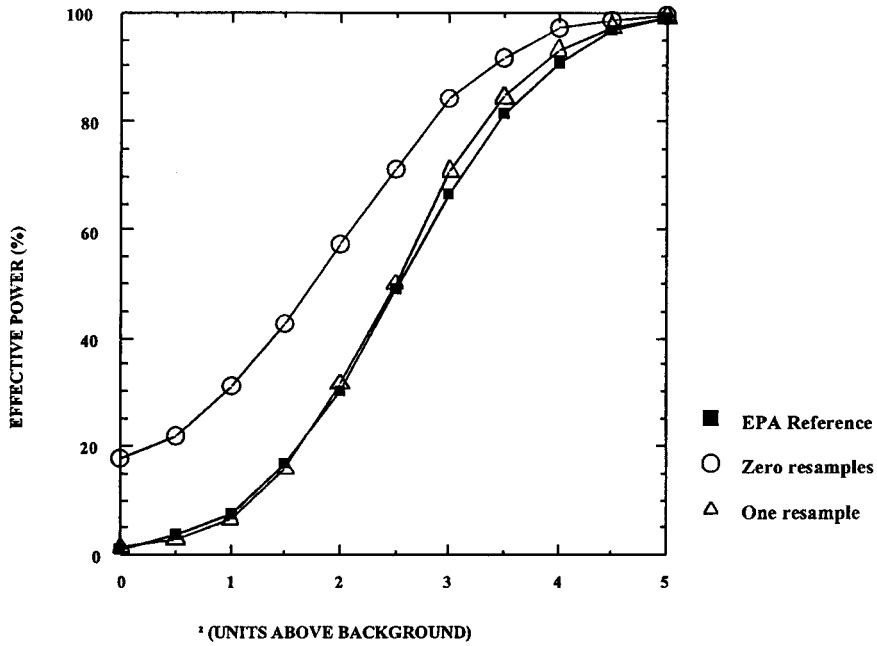



(16 Background Samples; 5 wells)
EFFECTIVE POWER (%)
100
80
60
40
20
0
0
1
2
3
4
5
² (UNITS ABOVE BACKGROUND)
EPA Reference
Two resamples (A)
Two resamples (B)

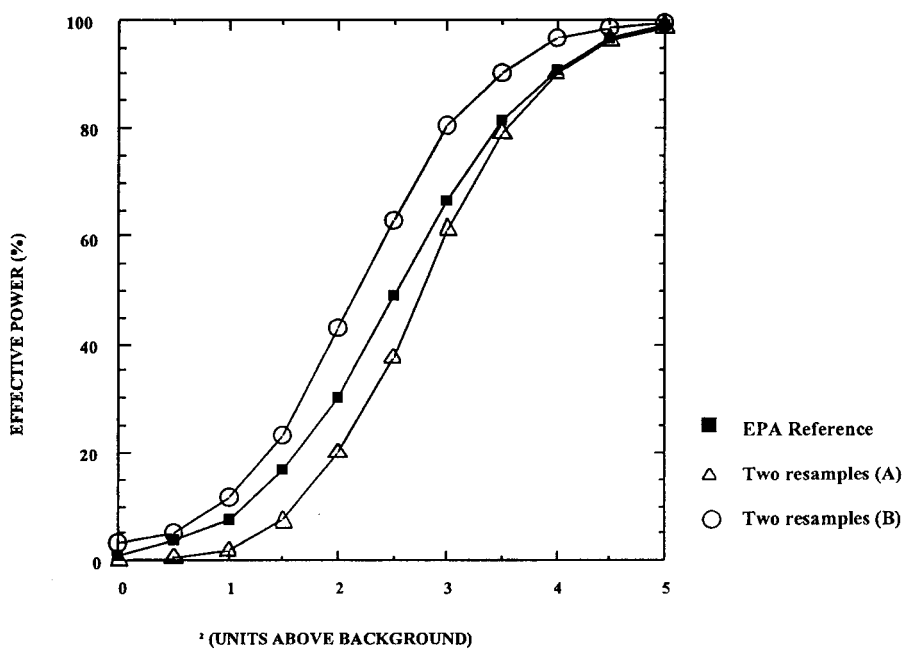

(16 Background Samples; 5 wells)
EFFECTIVE POWER (%)
100
80
60
40
20
0
0
1
2
3
4
5
² (UNITS ABOVE BACKGROUND)
EPA Reference
Three resamples (A)
Three resamples (B)

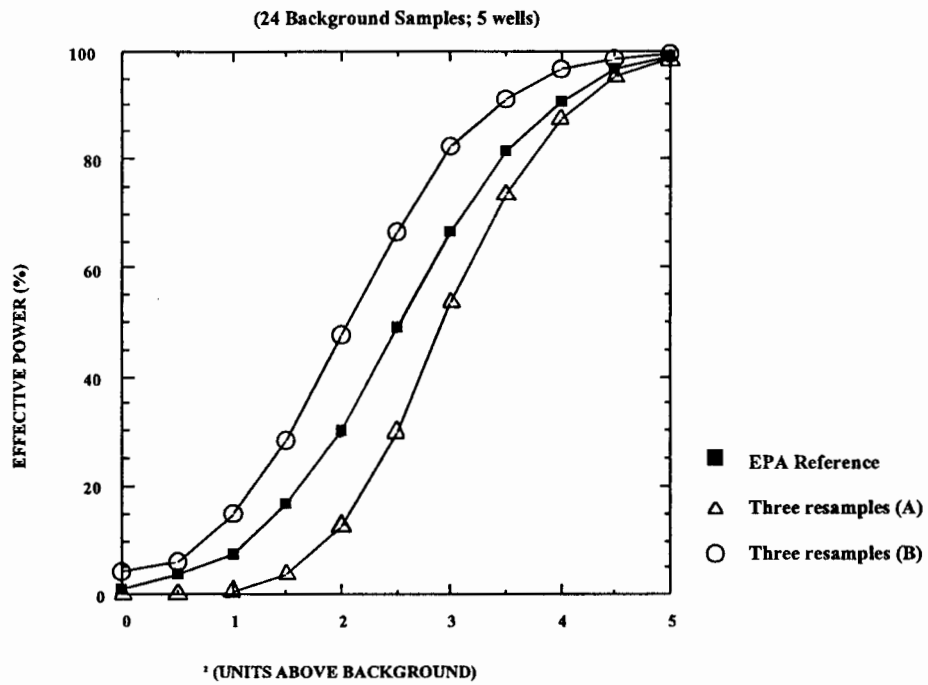
POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

(24 Background Samples; 5 wells)

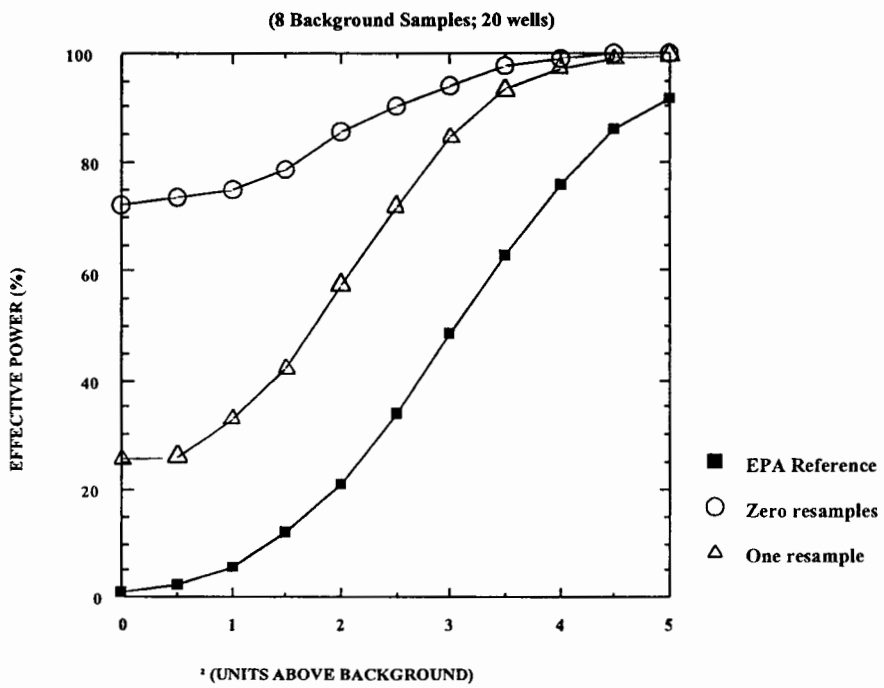


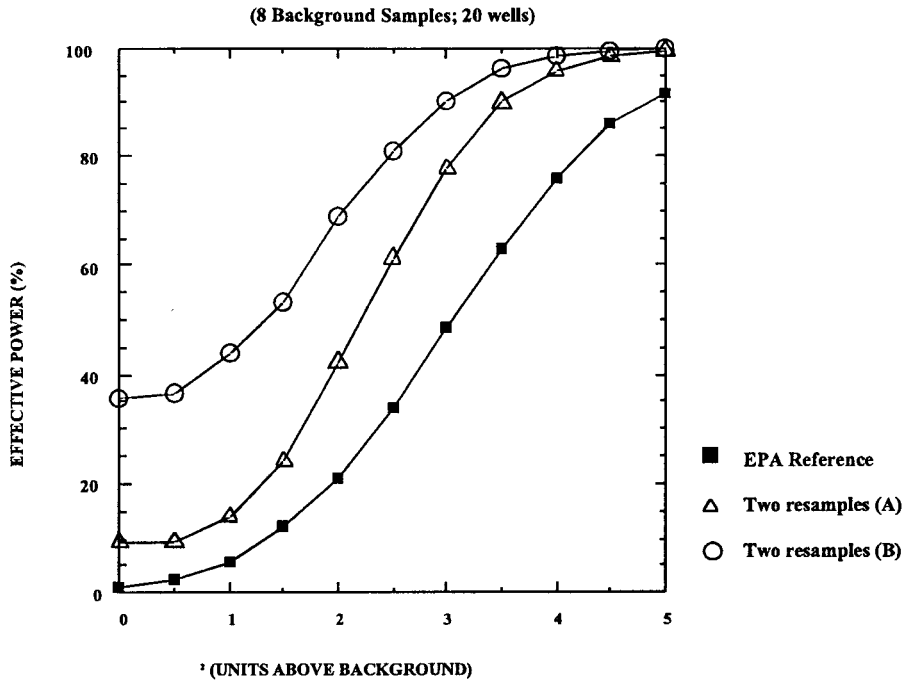
(24 Background Samples; 5 wells)

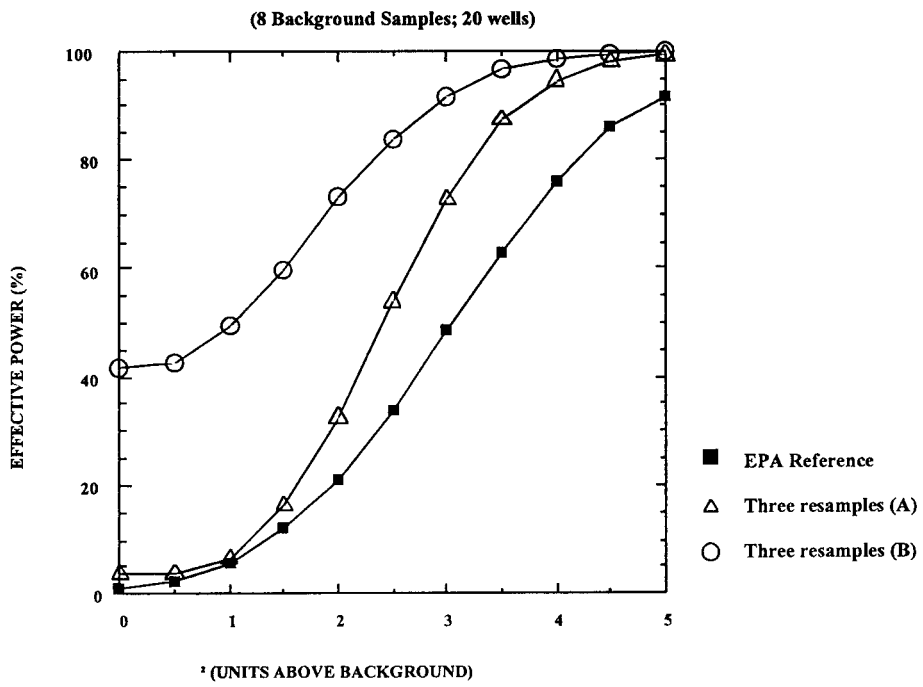




POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

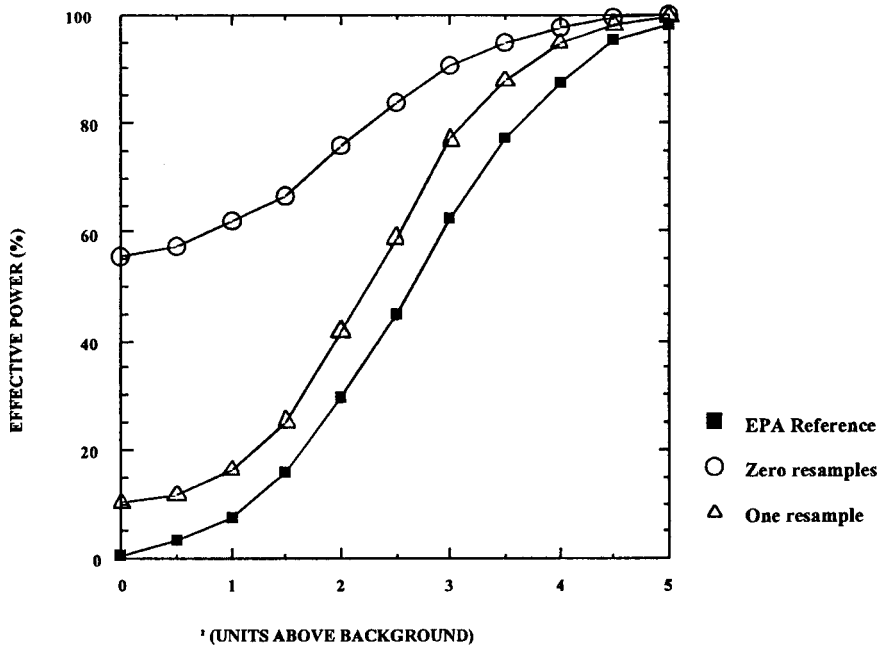



(8 Background Samples; 20 wells)
EFFECTIVE POWER (%)
100
80
60
40
20
0
0
1
2
3
4
5
² (UNITS ABOVE BACKGROUND)
EPA Reference
Two resamples (A)
Two resamples (B)

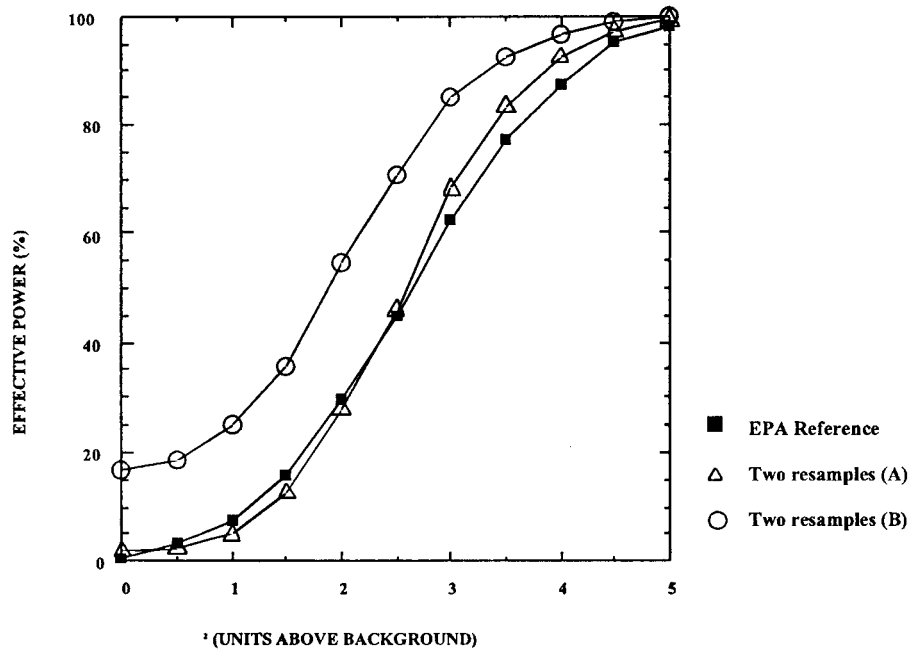

(8 Background Samples; 20 wells)
EFFECTIVE POWER (%)
100
80
60
40
20
0
0
1
2
3
4
5
² (UNITS ABOVE BACKGROUND)
EPA Reference
Three resamples (A)
Three resamples (B)

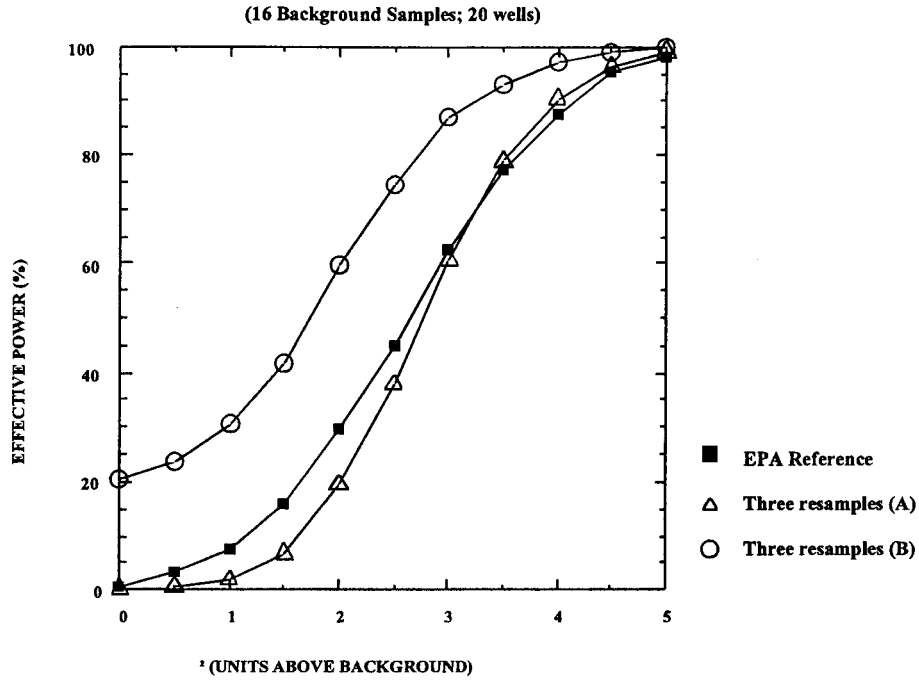
POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

(16 Background Samples; 20 wells)

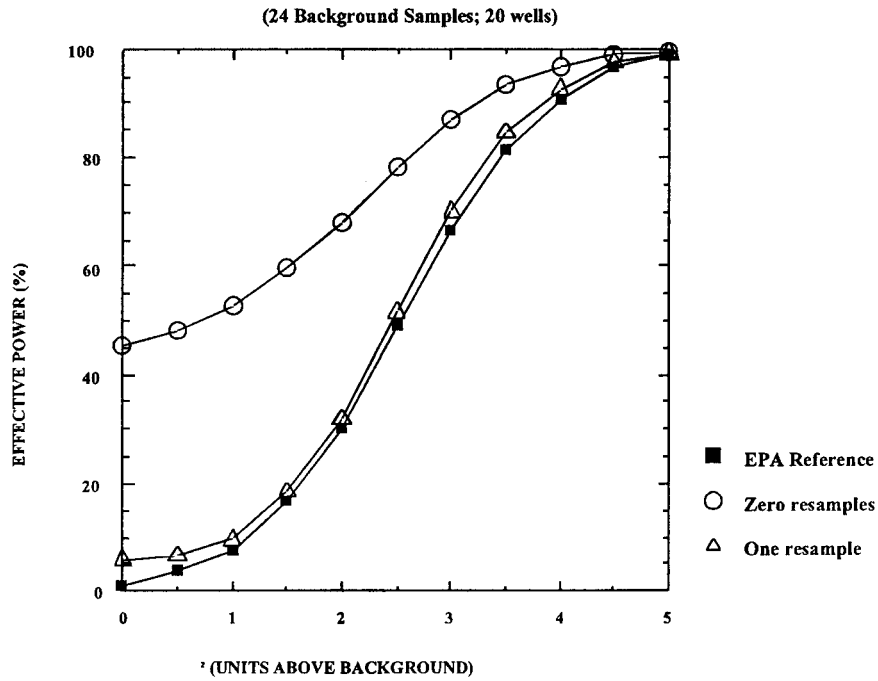


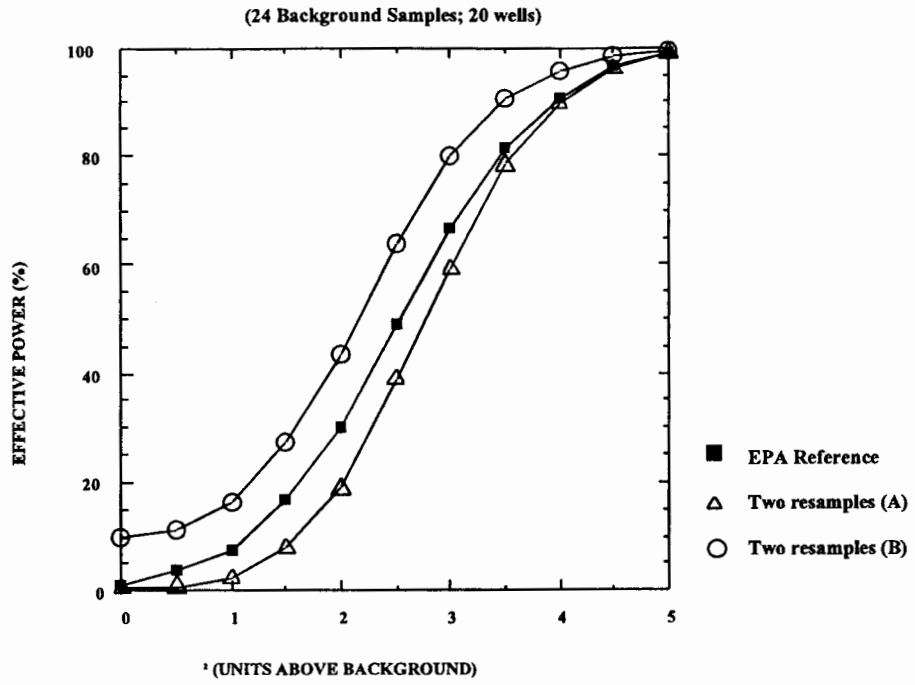
(16 Background Samples; 20 wells)

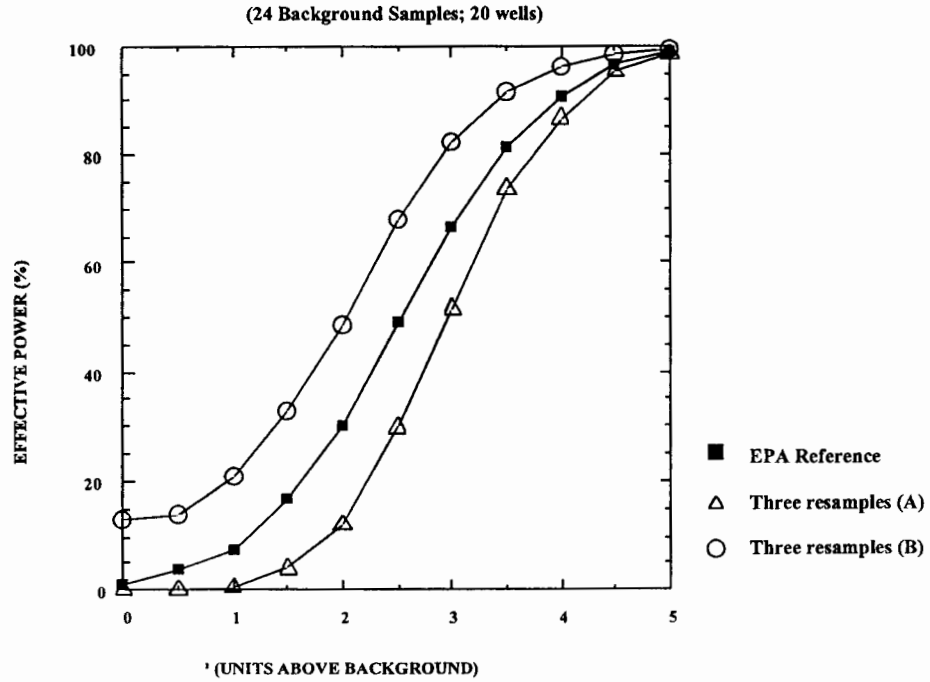




POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

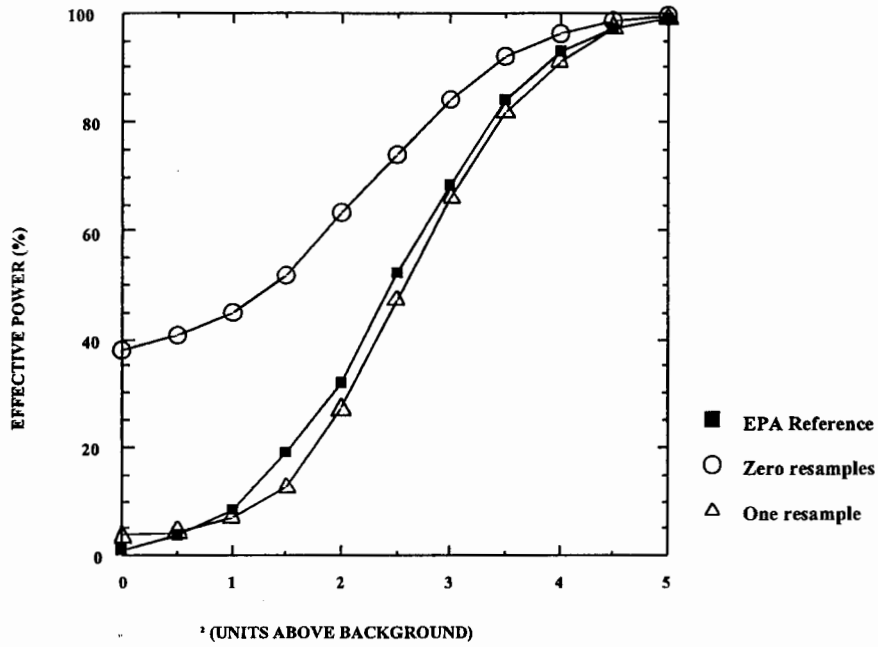



(24 Background Samples; 20 wells)
EFFECTIVE POWER (%)
100
80
60
40
20
0
0
1
2
3
4
5
(UNITS ABOVE BACKGROUND)
EPA Reference
Two resamples (A)
Two resamples (B)

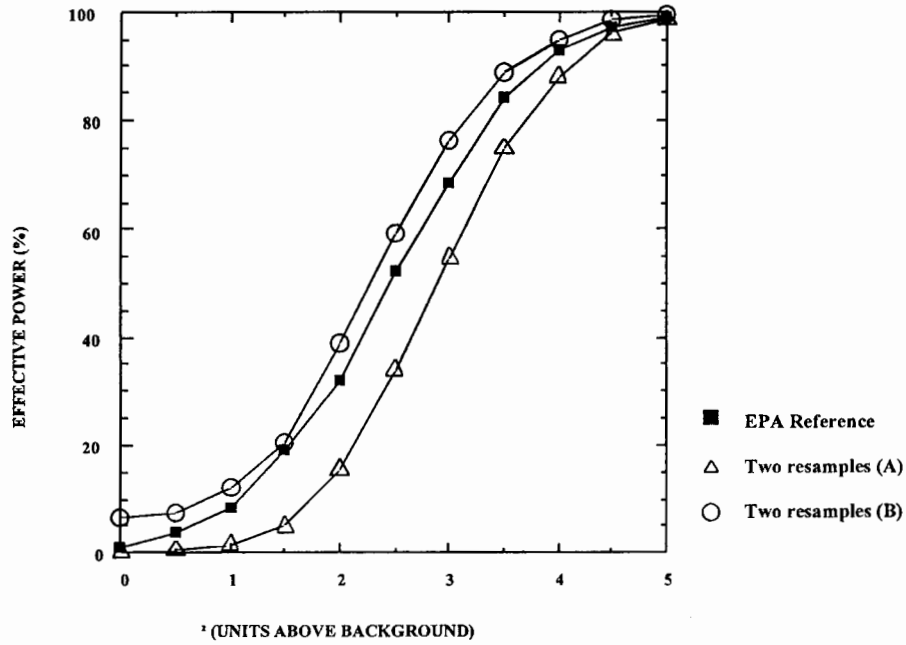

(24 Background Samples; 20 wells)
EFFECTIVE POWER (%)
100
80
60
40
20
0
0
1
2
3
4
5
(UNITS ABOVE BACKGROUND)
EPA Reference
Three resamples (A)
Three resamples (B)

POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

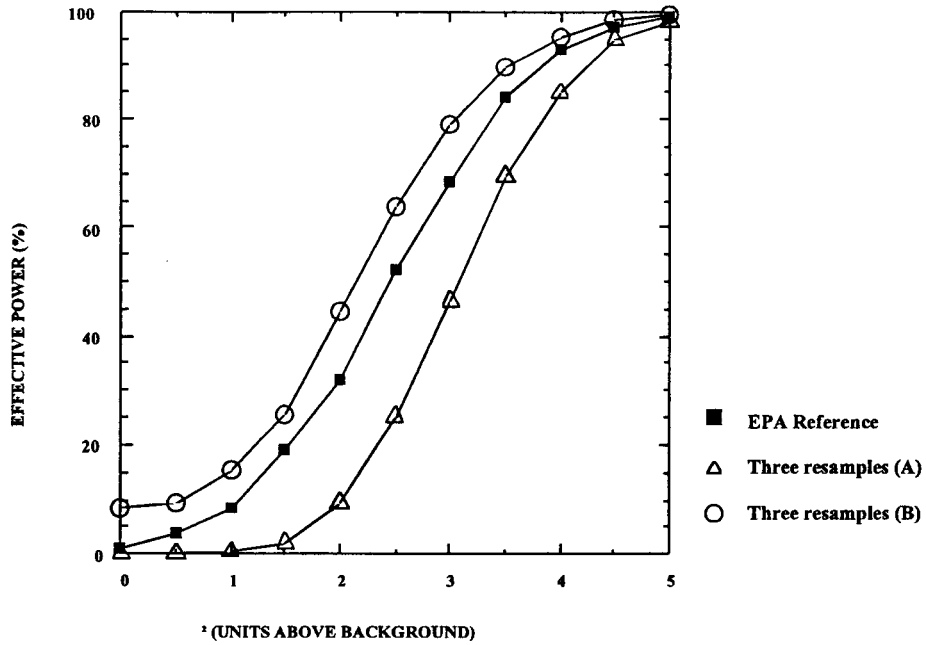
(32 Background Samples; 20 wells)



(32 Background Samples; 20 wells)

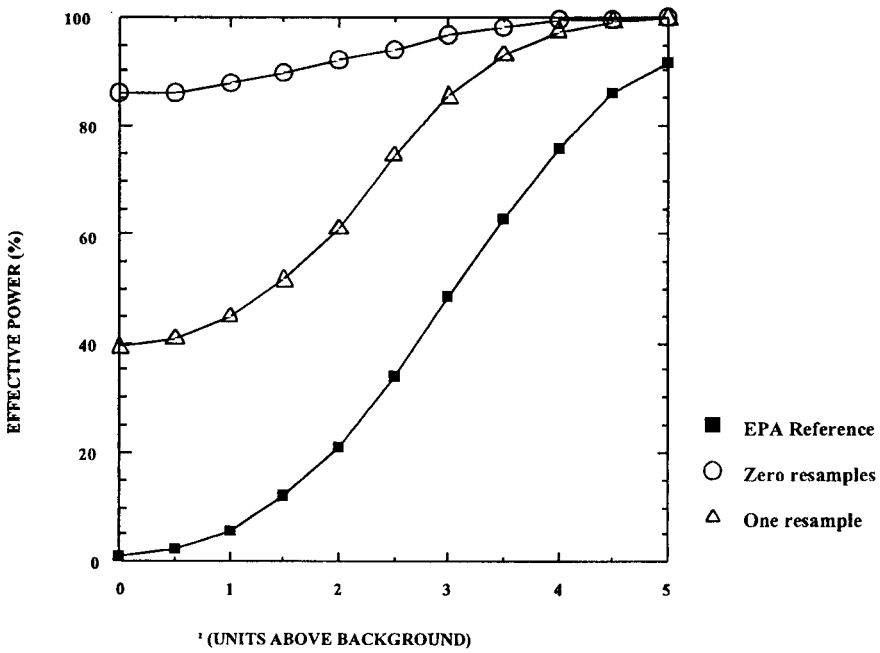


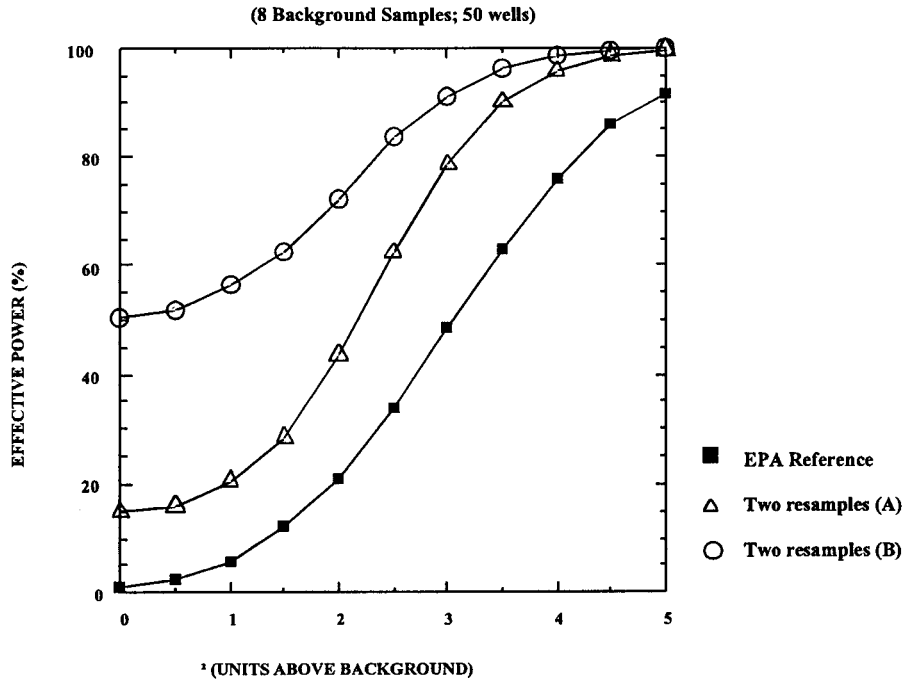
(32 Background Samples; 20 wells)

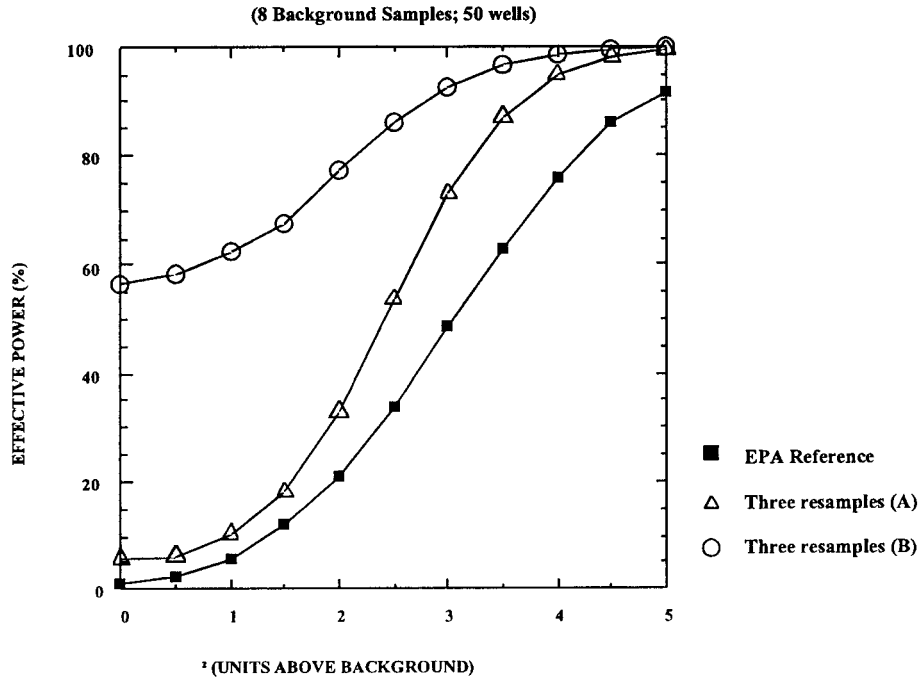


POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

(8 Background Samples; 50 wells)

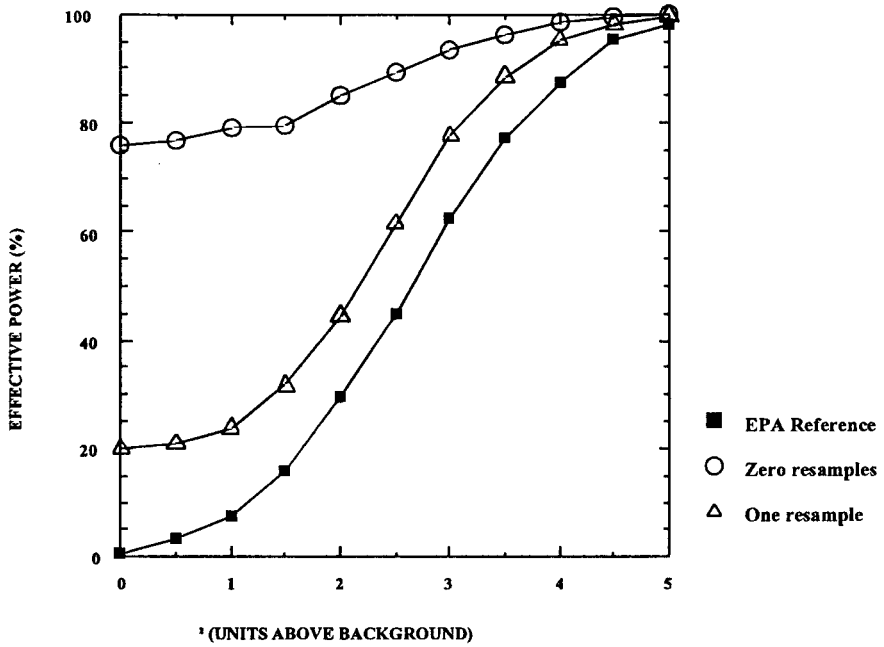



(8 Background Samples; 50 wells)
EFFECTIVE POWER (%)
100
80
60
40
20
0
0
1
2
3
4
5
² (UNITS ABOVE BACKGROUND)
EPA Reference
Two resamples (A)
Two resamples (B)

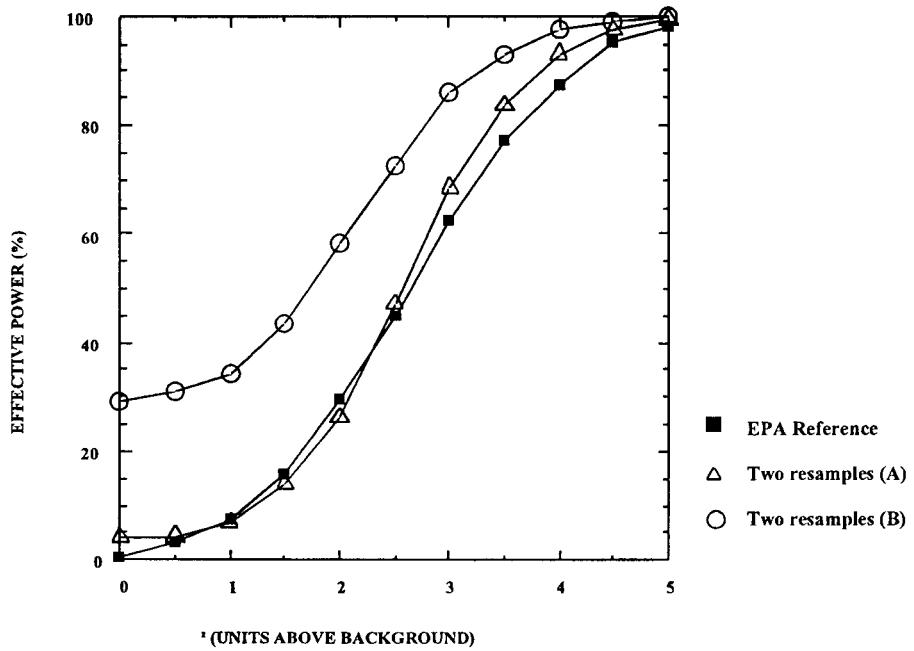

(8 Background Samples; 50 wells)
EFFECTIVE POWER (%)
100
80
60
40
20
0
0
1
2
3
4
5
² (UNITS ABOVE BACKGROUND)
EPA Reference
Three resamples (A)
Three resamples (B)

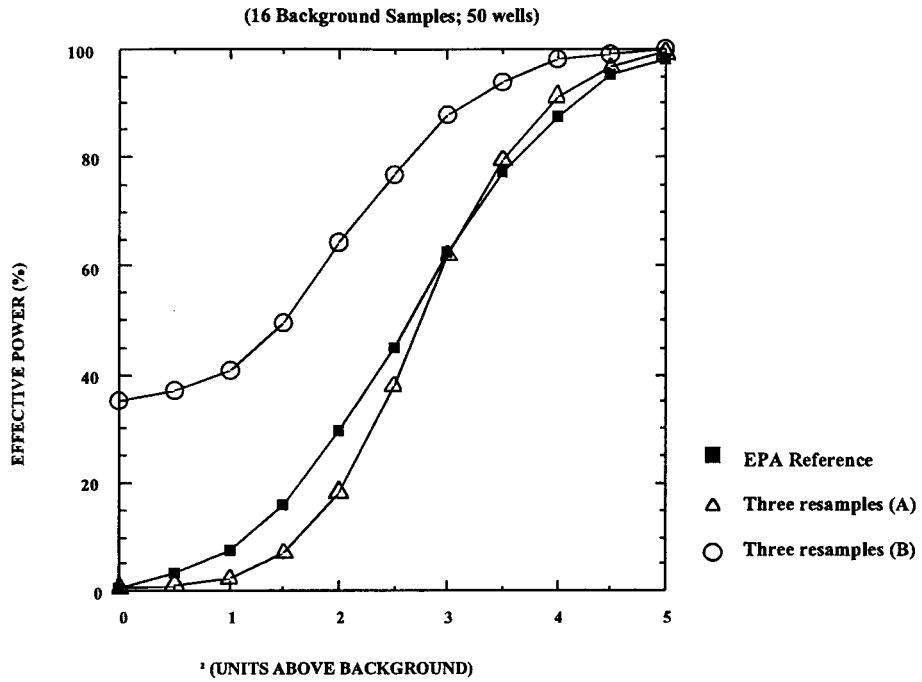
POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

(16 Background Samples; 50 wells)

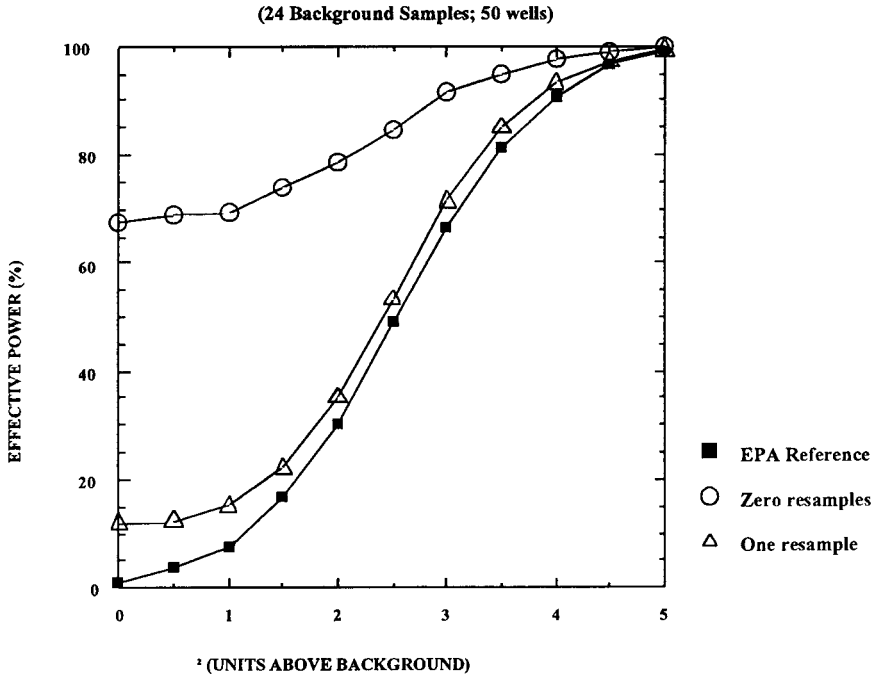


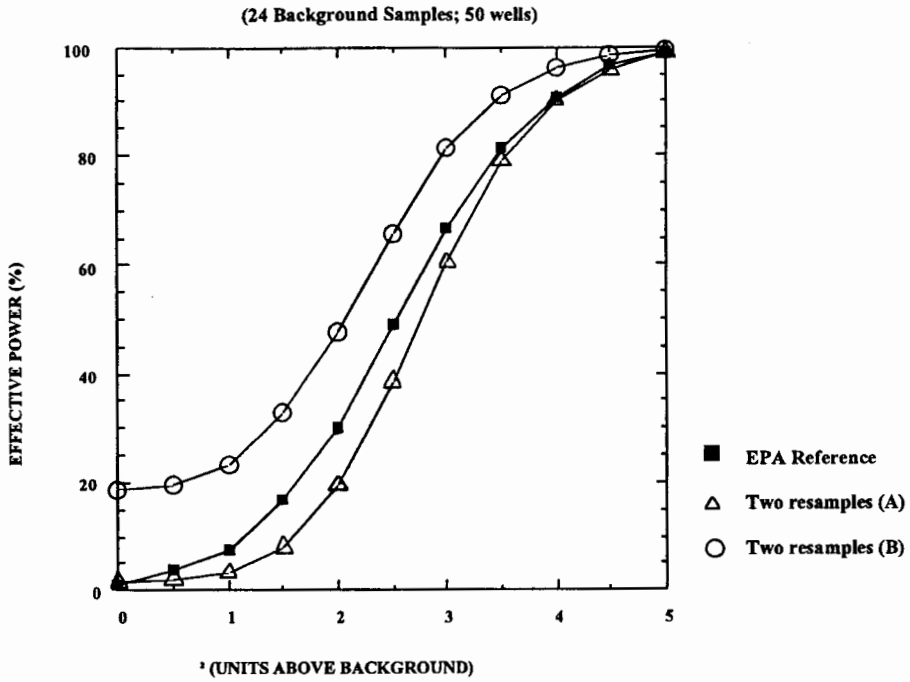
(16 Background Samples; 50 wells)

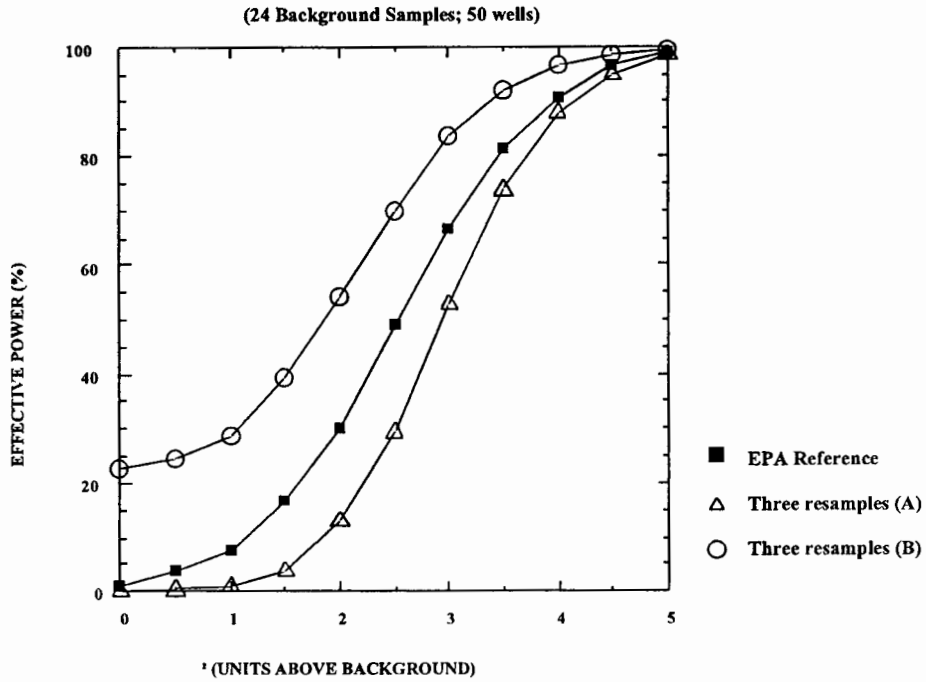




POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

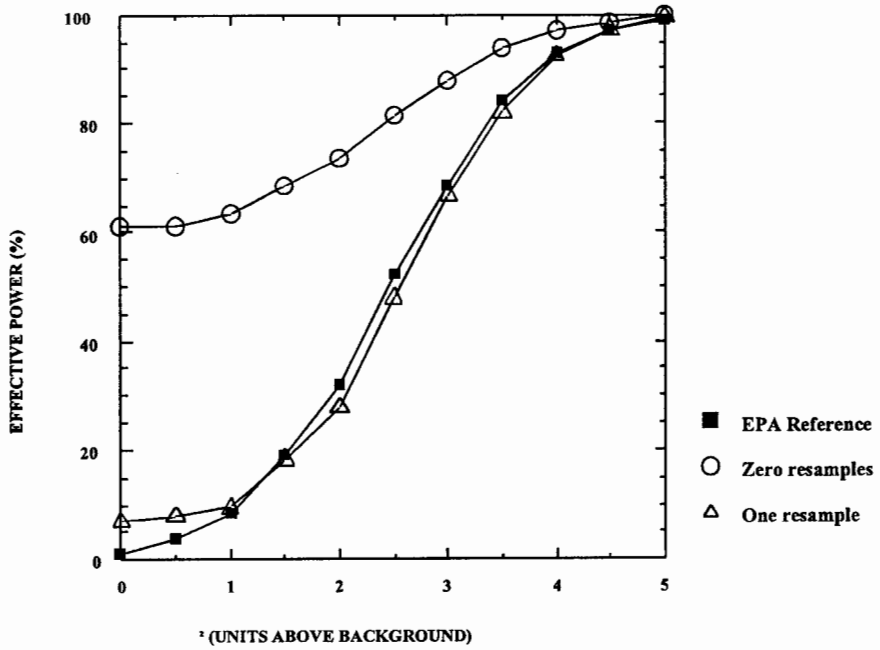



(24 Background Samples; 50 wells)
EFFECTIVE POWER (%)
100
80
60
40
20
0
0
1
2
3
4
5
² (UNITS ABOVE BACKGROUND)
EPA Reference
Two resamples (A)
Two resamples (B)

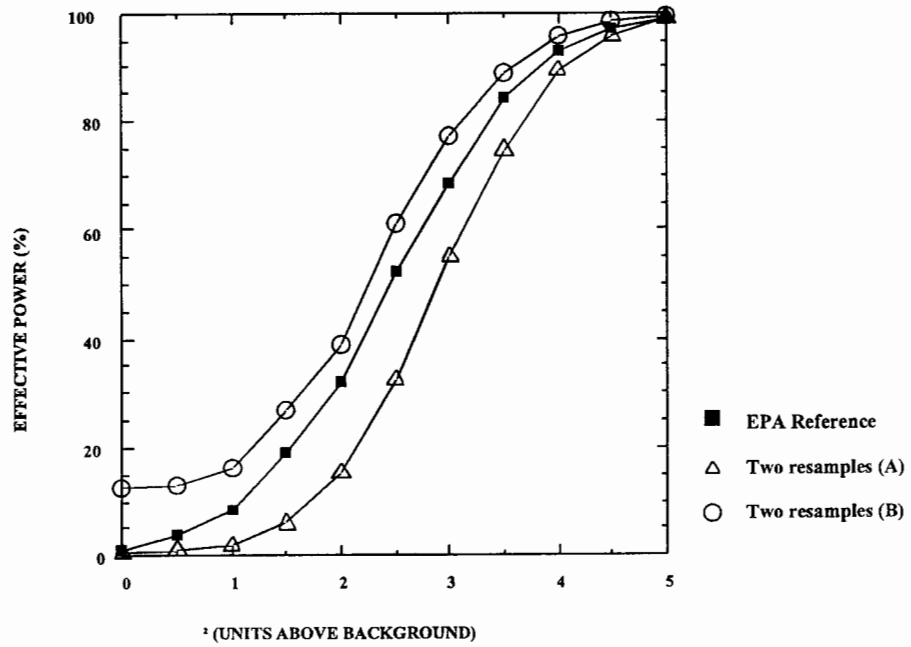

(24 Background Samples; 50 wells)
EFFECTIVE POWER (%)
100
80
60
40
20
0
0
1
2
3
4
5
² (UNITS ABOVE BACKGROUND)
EPA Reference
Three resamples (A)
Three resamples (B)

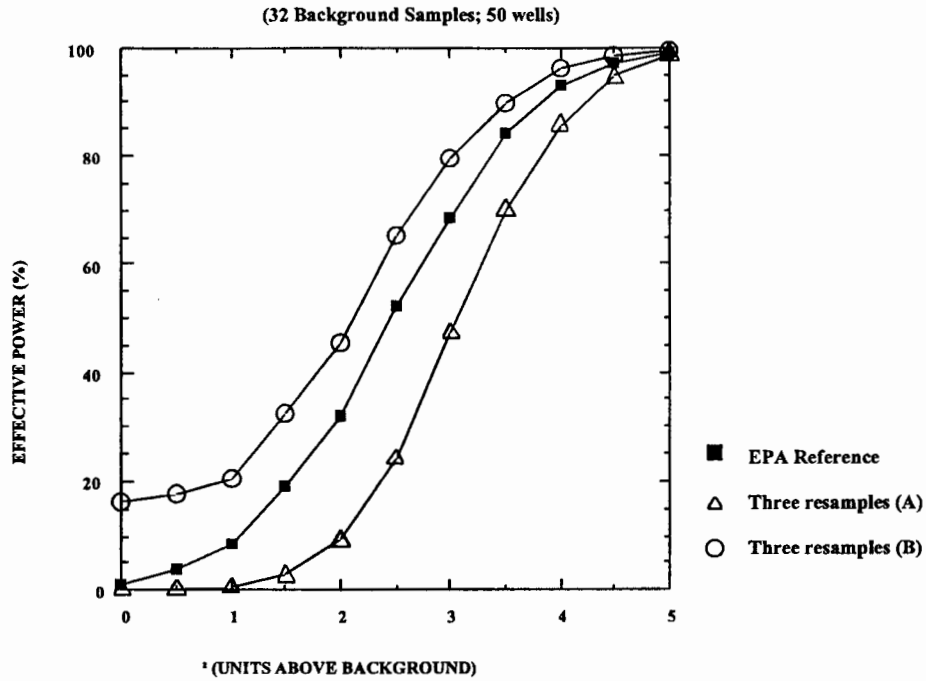
POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

(32 Background Samples; 50 wells)

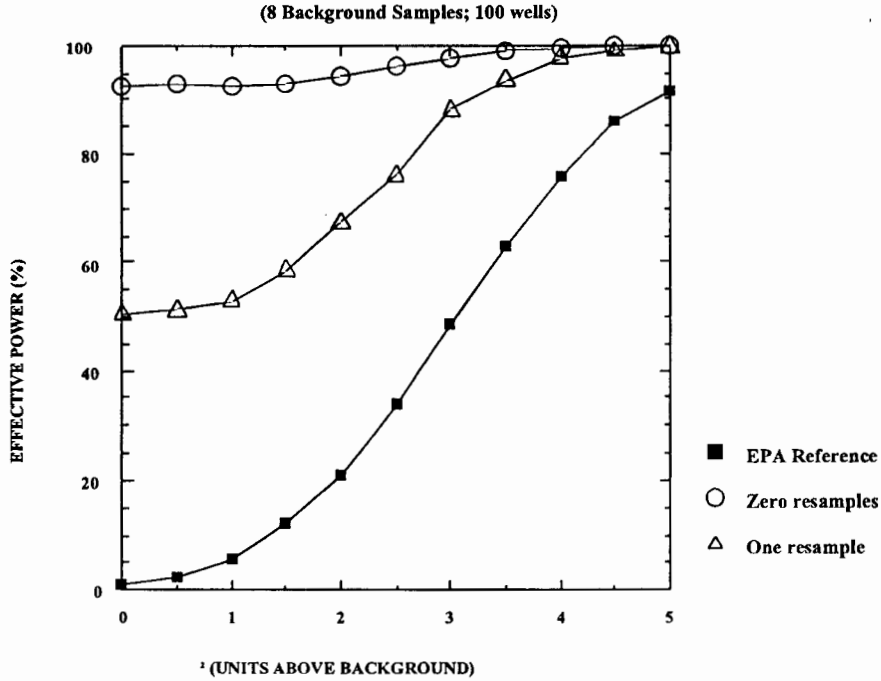


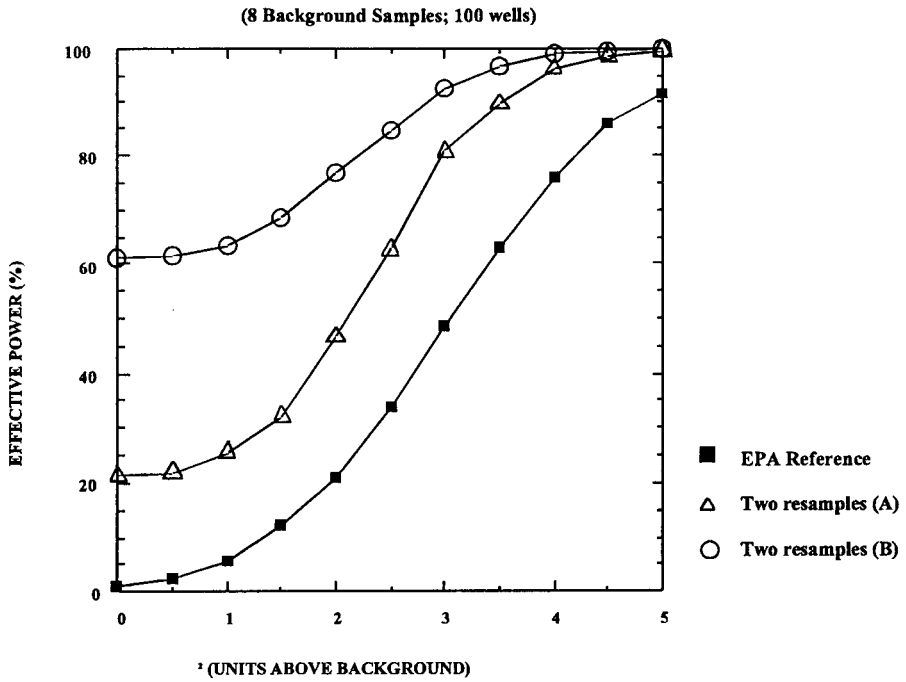
(32 Background Samples; 50 wells)

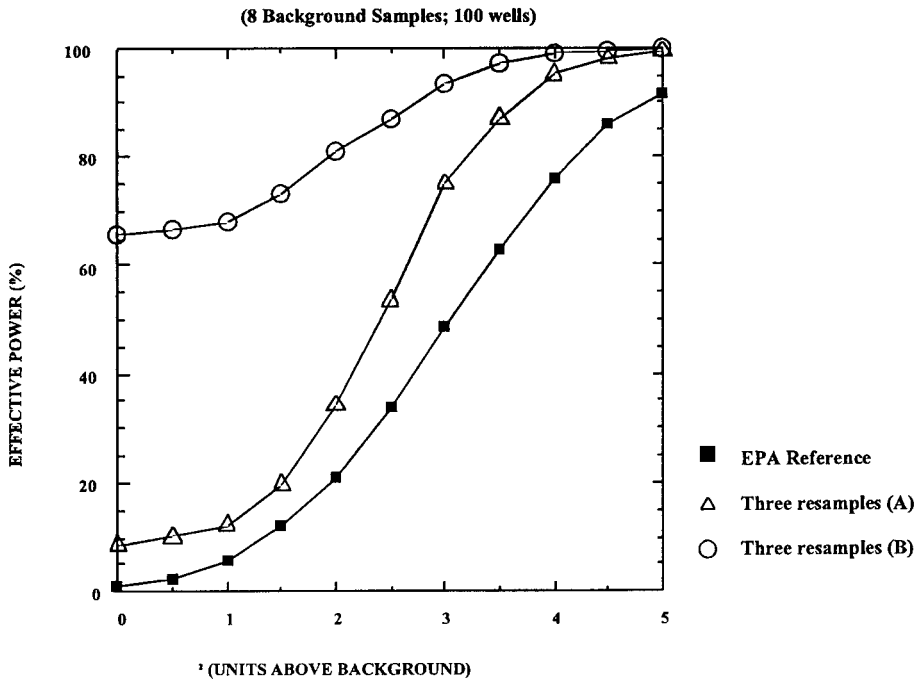




POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

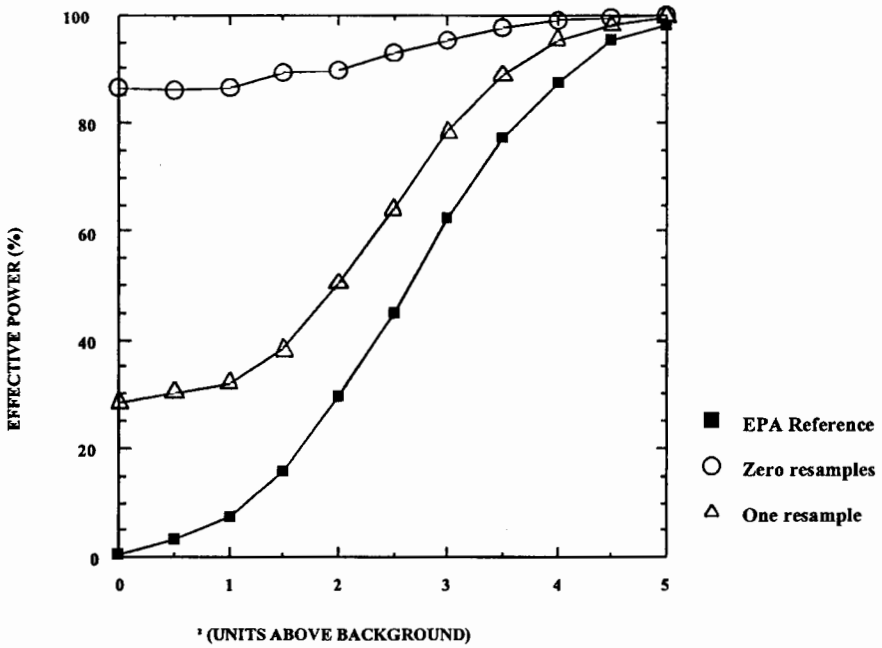



(8 Background Samples; 100 wells)
EFFECTIVE POWER (%)
100
80
60
40
20
0
0
1
2
3
4
5
² (UNITS ABOVE BACKGROUND)
EPA Reference
Two resamples (A)
Two resamples (B)

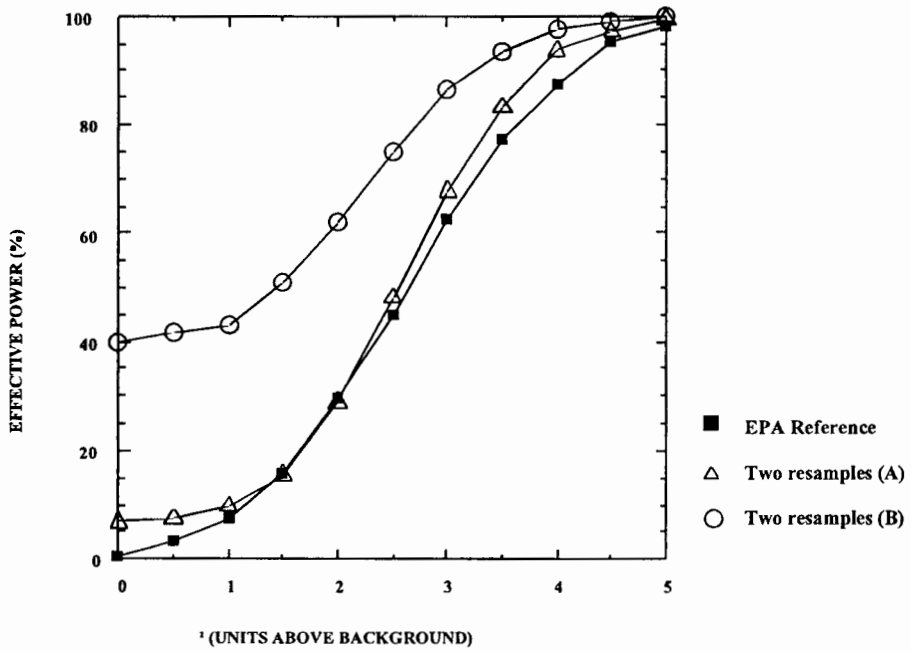

(8 Background Samples; 100 wells)
EFFECTIVE POWER (%)
100
80
60
40
20
0
0
1
2
3
4
5
² (UNITS ABOVE BACKGROUND)
EPA Reference
Three resamples (A)
Three resamples (B)

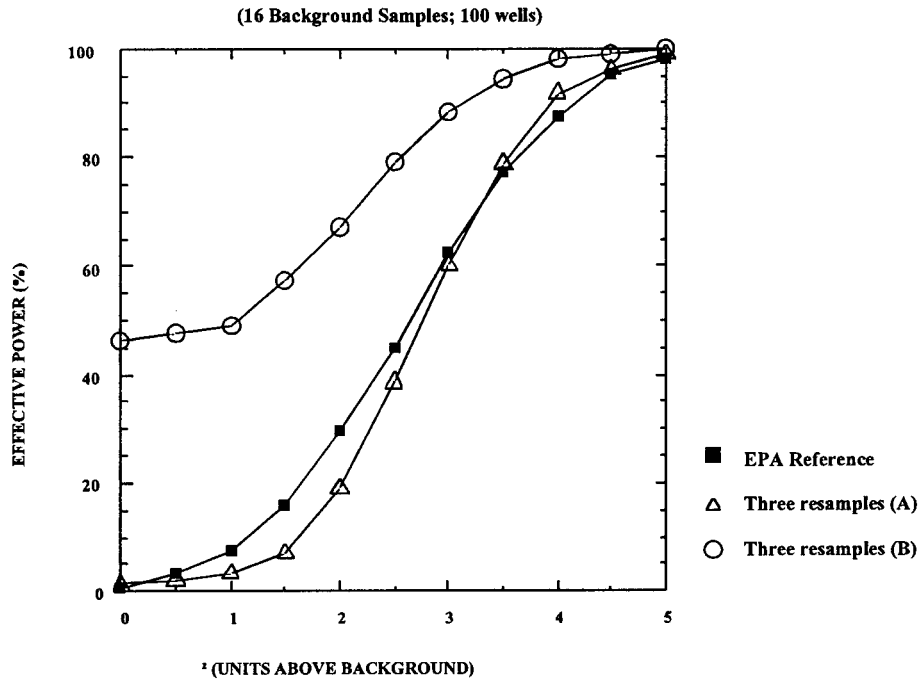
POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

(16 Background Samples; 100 wells)

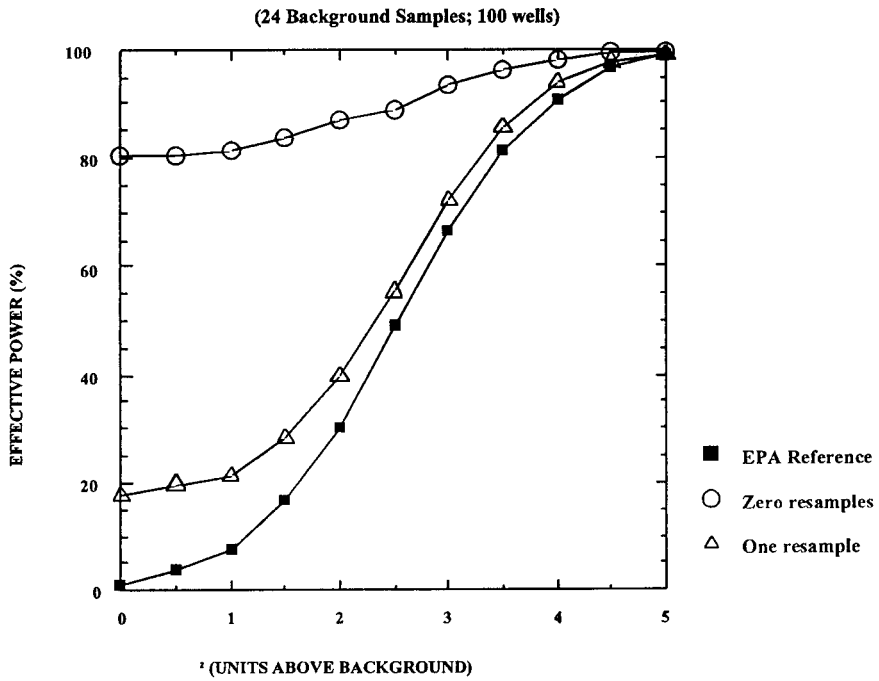


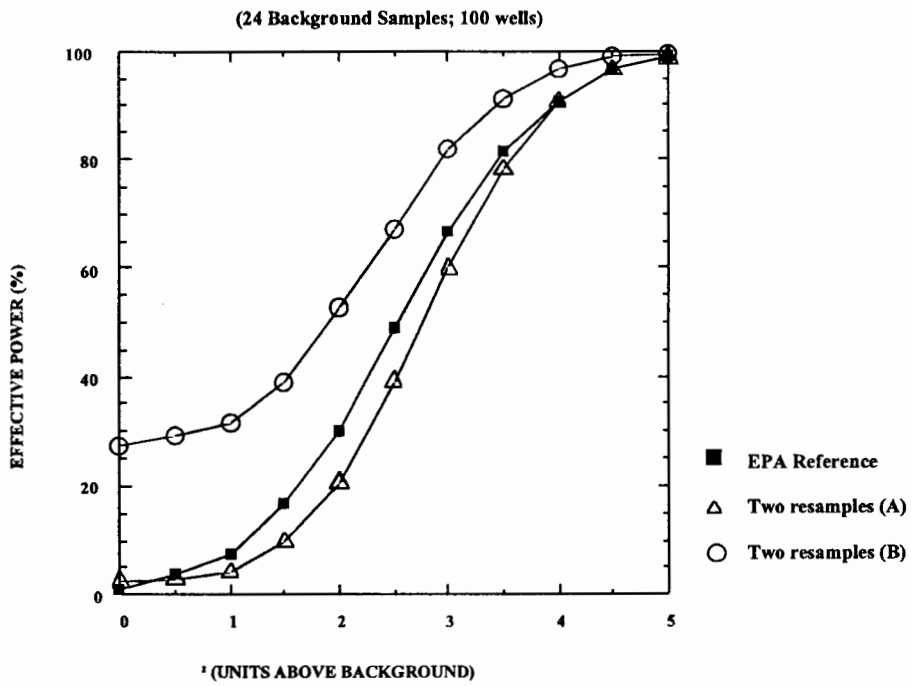
(16 Background Samples; 100 wells)

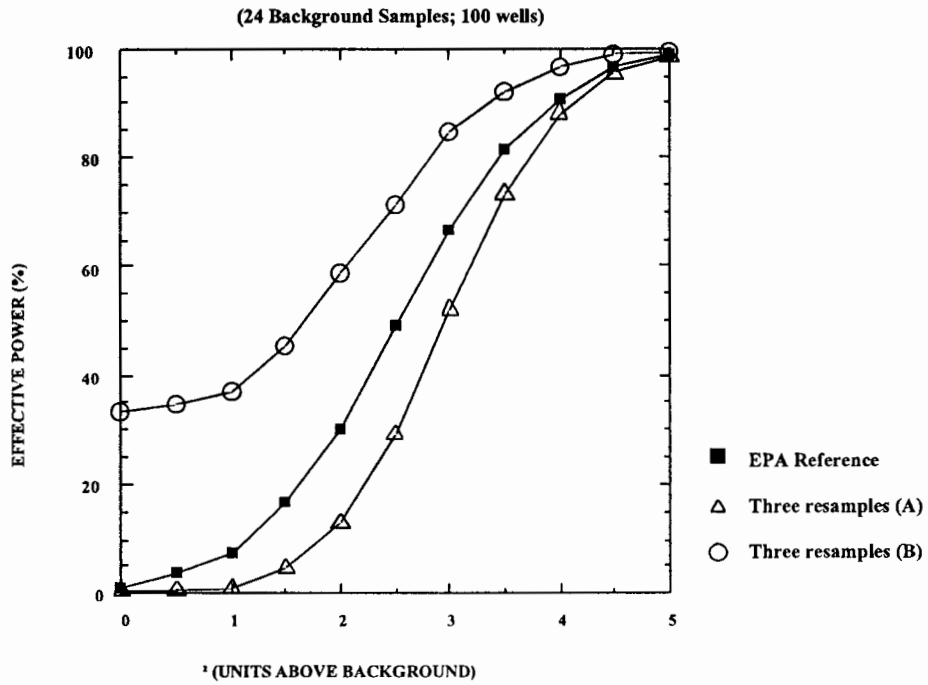




POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

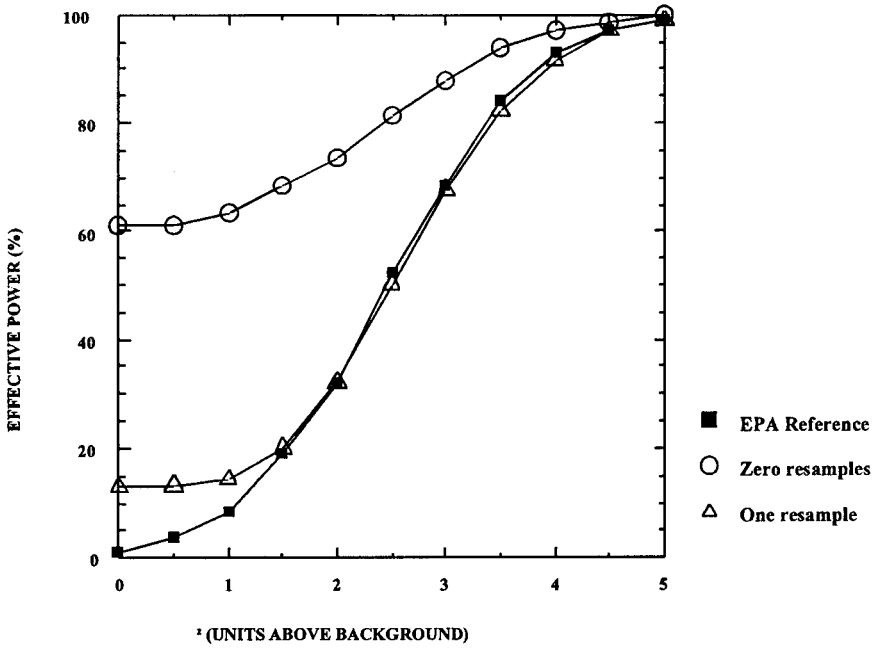



(24 Background Samples; 100 wells)
EFFECTIVE POWER (%)
100
80
60
40
20
0
0
1
2
3
4
5
(UNITS ABOVE BACKGROUND)
EPA Reference
Two resamples (A)
Two resamples (B)

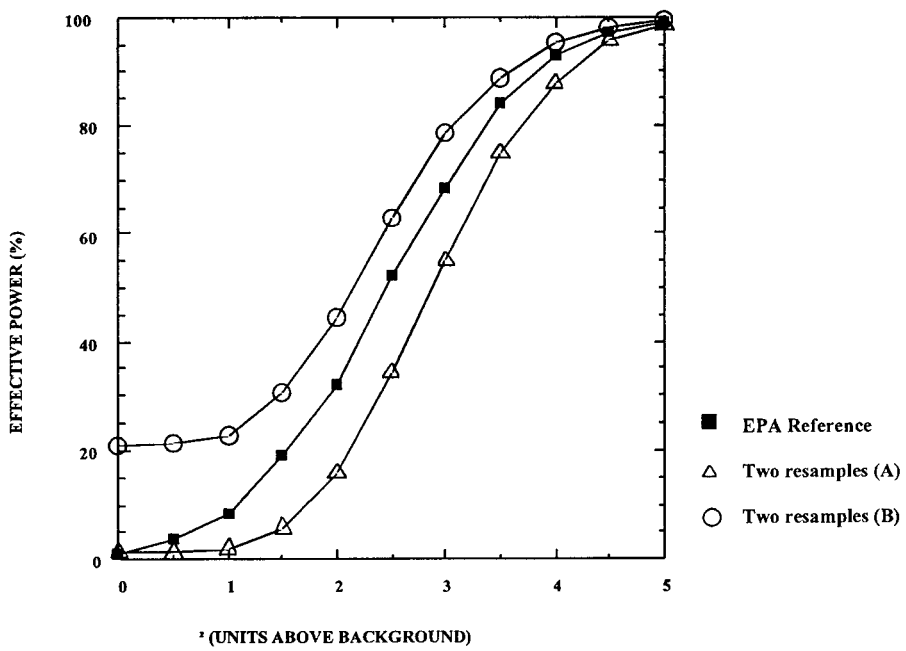

(24 Background Samples; 100 wells)
EFFECTIVE POWER (%)
100
80
60
40
20
0
0
1
2
3
4
5
(UNITS ABOVE BACKGROUND)
EPA Reference
Three resamples (A)
Three resamples (B)

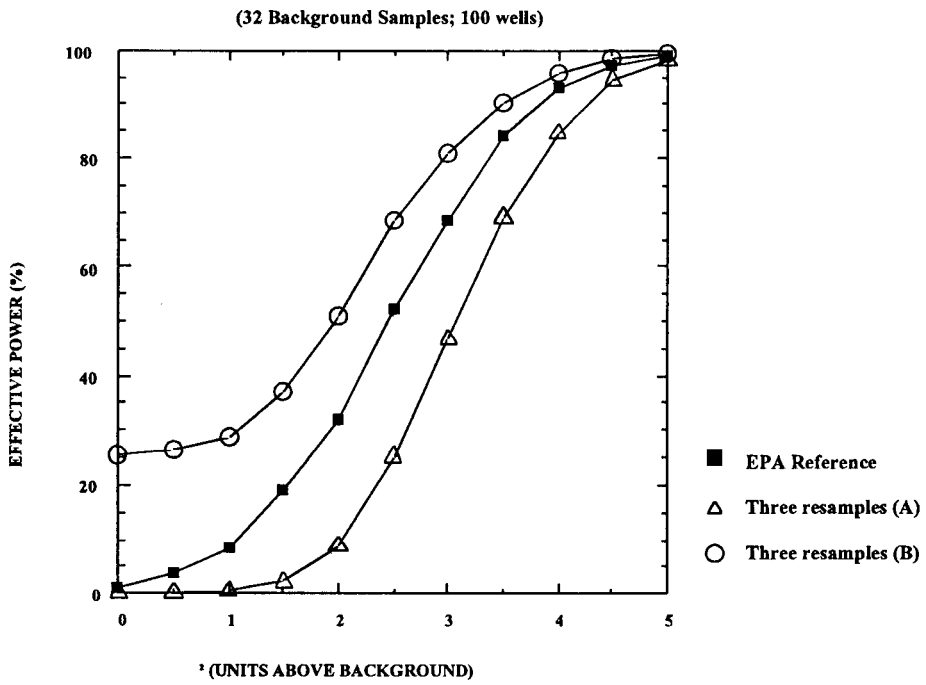
POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

(32 Background Samples; 100 wells)



(32 Background Samples; 100 wells)




(32 Background Samples; 100 wells)
EFFECTIVE POWER (%)
100
80
60
40
20
0
0
1
2
3
4
5
² (UNITS ABOVE BACKGROUND)
EPA Reference
Three resamples (A)
Three resamples (B)