530R93003

Statistical Training Course

for

Ground-Water Monitoring Data Analysis

Sponsored by the

U.S. Environmental Protection Agency
Office of Solid Waste

1992

OUTLINE OF STATISTICAL TRAINING COURSE

August 30, 1994

I. COURSE INTRODUCTION

A. Objectives

1. Review regulations pertaining to the statistical evaluation of ground-water monitoring data
   a. Provide layman's guide to the regulatory requirements
2. Introduce GRITS/STAT Statistical Software
3. Provide an intuitive understanding of statistical thinking and analysis
   a. Expand one's statistical vocabulary
   b. Learn the right questions to ask when analyzing data
   c. Basic distributional models
4. Discuss basic techniques for interval estimation and hypothesis testing
   a. ANOVA, t-tests
   b. Confidence, Tolerance, and Prediction Intervals
   c. Control Charts
   d. Parametric versus non-parametric procedures
5. Learn specific statistical procedures for analyzing ground-water monitoring data:
   a. Indicate appropriate use and how to use
   b. Provide warnings, limitations, and interpretations of results
   c. Discuss impact of specific assumptions on method performance

B. Summary of Current Regulations/Guidance

1. Statistical Analysis of Ground-Water Monitoring Data Analysis at RCRA Facilities: Final Rules
   a. Subtitle C regulation: 53 FR 39720; October 11, 1988 (Hazardous Wastes)
   b. Subtitle D regulation: 56 FR 50978; October 9, 1991 (Municipal Solid Wastes)
2. Interim Final Technical Guidance Document: April 1989
3. Addendum to Interim Final Guidance: July 1992

4. Older regulations only listed CABF Student's t-test

5. Regulations applied CABF t-test incorrectly

a. Replicate sampling procedure

i. Assumes static background
ii. Replicate samples not independent
iii. Too many false positives due to unreasonably low estimated variance

b. Assumes normal distribution of original data

i. Groundwater data often lognormal
ii. Did not handle nondetects appropriately

c. Could not accomodate large number of comparisons

i. High false positive rates

d. CABF t-test not a bad procedure but only appropriate for certain cases

6. "New" statistical procedures

a. Give more flexibility and adaptability to specific groundwater environments

b. Sampling procedures based on site hydrogeology, not lab replicates

c. Can accomodate:

i. Departures from normality
ii. Unequal variances
iii. Temporal and spatial variability
iv. Nondetects

d. Methods include:

i. Parametric and non-parametric ANOVA
ii. Parametric and non-parametric t-tests
iii. Confidence, Tolerance, and Prediction Intervals
iv. Control charts
v. Alternative procedures

C. Overview of Statistical Requirements

1. Statistical tests required for each monitored constituent in each downgradient well

2. Statistical method must be chosen from a list of options and specified within the operating record

3. The statistical method must comply with certain performance standards

a. Must be appropriate to observed distribution of data

b. Must meet minimum false positive rates

c. Parameters for specific procedures must be protective of human health and the environment

4. Unless comparing compliance data to a regulatory standard (e.g., MCL), statistical comparison must be made between downgradient and background measurements

   a. Possible release indicated if there is a statistically significant increase over background

   b. Exceptions/alternatives for other cases detailed in guidance documents

II. LAYMAN'S GUIDE TO THE STATISTICAL REQUIREMENTS

A. Tests required for each constituent in each well

1. Rationale: if a release should occur, want to identify with some certainty which specific constituent is polluting which specific well

2. Regulations prohibit the "pooling" of constituents in designing statistical procedures

   a. Pooling of constituents would involve testing the results from more than one constituent simultaneously in an "omnibus-type test" such as ANOVA

   b. Danger: Omnibus tests require similar distributions in each of the data sets being pooled

      i. Distinct constituents can have very different observed distributions
      ii. Can lead to misleading test results

   c. Danger: The number of constituents for which testing must be performed can be very large

      i. Using an omnibus test to combine a large number of "clean" parameters and one "dirty" parameter can lead to misleading results
      ii. Release indicated by the "dirty" parameter can be overlooked because the test shows no overall statistical difference

3. Regulations also prohibit "pooling" of wells

   a. Two different types of "pooling" when it comes to wells; one type OK, the other not

   b. Inappropriate pooling of wells would involve testing of downgradient measurements in one group as if the well identifiers had been discarded

      i. Example: using t-test to compare background against the pooled measurements from two downgradient compliance points

   c. Danger: if a significant statistical result is obtained, which compliance point is contaminated? Difficult to say and certainly not indicated by the statistical results

   d. Appropriate pooling of wells involves testing multiple wells simultaneously for a single constituent with an "omnibus" test, where the well identifiers are kept intact and are built into the testing procedure

i. "Omnibus" procedures like ANOVA keep the data at each well separately identified, so that if an overall difference is found, the individual contaminated well can be identified on the basis of the statistical results and further testing

ii. Example: ANOVA would be used on the first example in two steps; if the overall test showed a significant difference between the compliance points and background, additional calculations would be made to identify the contaminated well

B. Tests must be chosen from a particular list of options

1. Rationale: given the wide variety of statistical tests in existence and the vastly different assumptions and requirements associated with these tests, EPA has tried to provide a reasonable set of alternative tests that find practical application to groundwater data

2. EPA recognizes the need to standardize the set of potential tests down to a reasonable few, for the sake of consistent evaluation by different analysts, and yet to allow flexible adaptation of statistical testing strategies to a wide variety of monitoring scenarios and observed data

   a. Regulations allow for alternative tests not explicitly listed, if the test can be shown to be applicable and to meet relevant performance criteria

3. The set of "standard" statistical options includes: ANOVA (parametric and non-parametric versions), t-tests, Control Charts, and Statistical Intervals (confidence, prediction, and tolerance intervals)

   a. t-tests and confidence intervals are not explicitly mentioned in the current regulations, but have been approved for use by EPA on a regular basis

4. ANOVA: Analysis of Variance

   a. Use to compare background data versus measurements from one or more downgradient wells

   b. Two-step procedure

      i. Overall test on all the data identifies any possible statistical differences
      ii. If overall ANOVA significant, individual contrasts are run to compare background data versus each individual downgradient well

   c. In parametric version, original or log-transformed measurements are used; test evaluates whether the mean concentration levels from any two groups being tested are significantly different

   d. In nonparametric setting, test is based on ranks of data rather than measurements themselves

      i. Test known as Kruskal-Wallis procedure
      ii. Test evaluates whether the median concentration levels from any two groups are significantly different

5. t-tests

   a. More or less an ANOVA procedure run on only two groups: one set of background data and measurements from one downgradient well

   b. CABF t-test is one of many types of t-tests

6. Control Charts

   a. Involves data from a single well, plotted over time on a special graph

   b. Well must be initially clean if used for detection monitoring; background information can be collected at this well or from other similar background locations

   c. Method allows visual tracking of constituent behavior at the well over time and visual identification of possible contamination

   d. Good for intrawell comparisons, when measurements cannot be directly compared with data from other background locations (perhaps due to heavy on-site spatial variability)

7. Statistical Intervals: Confidence, Prediction, Tolerance

   a. These methods often used for special circumstances

   b. Use confidence or tolerance intervals when comparing downgradient measurements from a well against a known regulatory standard (e.g., MCL)

   c. Use prediction intervals when doing intrawell comparisons or when comparing very limited compliance data versus background (e.g., collection of one compliance sample per well every 6 months)

   d. All interval procedures estimate a range of values designed to represent some aspect of either the background or downgradient well populations

C. Statistical method must comply with performance standards

1. Must be chosen in accordance with observed distribution of data

   a. All statistical tests assume something about the distribution of data

      i. Parametric tests assume the data follow a specific form like the normal or lognormal distribution
      ii. Nonparametric tests usually assume the data are symmetric or perhaps identically distributed from group to group

   b. Distributional assumptions can be critical to getting the right answer from a test

      i. Example of lognormal benzene data

   c. Care must be taken to match the statistical method with what is known about the data distribution

      i. Transform the data or change the method used if warranted by the data distribution

   d. Common examples:

      i. Normal versus lognormal data in parametric tests
      ii. Data with many non-detects

2. Test must meet minimum false positive rates

   a. Seems odd: if a false positive (i.e., identifying contamination at a "clean" well) is a bad thing, why force the statistical test to operate under a <u>minimum</u> false positive level?

b. Key is relationship between false positives, false negatives, and statistical power

i. Statistical power measures a test's ability to identify contamination when it in fact exists

ii. False negatives occur when test misses real contamination

c. Since statistical power is inversely related to false negative rate, increasing the power will lower the chance of false negatives and raise the efficiency of the test in finding contamination

e. But, lower false positive rates generally also linked with lower statistical power

f. To maintain certain level of power, must not allow false positive rate to drop too much; hence the regulation that individual comparisons have a false positive rate of at least 1% and ANOVA tests have a false positive rate of at least 5%

g. Overriding EPA goal: maintain adequate statistical power so that contamination is identified when it exists

3. Parameters must be protective of human health and the environment

a. For most statistical procedures, certain parameters or settings must be chosen to calibrate the method to a specific data application (e.g., false positive rate)

b. Choice is not arbitrary, but must be such that adequate levels of statistical power are maintained while at same time minimizing false positive rates to the extent possible

i. Regulations allow for flexibility in design parameters for prediction and tolerance intervals and control charts as long as "reasonable confidence" test is met

4. If necessary, test must appropriately account for nondetects

a. Often need to use a non-parametric procedure as an alternative

5. If necessary, test must control for temporal and spatial variability

a. May need to run an intrawell comparison or to compute spatially- or time-adjusted measurements

D. Notes on Establishing Background Data

1. Establish background data with intent of

a. Gauging average levels and variability in naturally-occurring constituents, or

b. Confirming the absence of other constituents

2. Beware the consequences of small background sample sizes

a. Much more statistical power comes from larger sample sizes

b. Statistical tests do not operate on a pass/fail basis, but rather on a no-decision/fail basis

i. Sometimes a statistical test will be inconclusive (i.e., not fail) simply because the sample size is too small

ii. Such cases could lead the analyst to miss possible evidence of contamination

3. Therefore, sample from background wells as often as is feasible

   a. Better to sample a few constituents frequently than many constituents infrequently

   b. Replicate samples are not statistically independent and do not count as separate samples

   c. When background wells cannot be sampled frequently, consider pooling data from multiple background wells to increase overall background sample size

4. When should data from multiple upgradient wells be pooled for statistical purposes?

   a. Wells should generally be screened in same hydrostratigraphic unit

   b. Groundwater chemistry should be similar

   c. Comparisons should be made with bar charts, pie charts, and tri-linear diagrams of major constituent ions

5. In intrawell comparisons, the historical or past data from the well is often treated as the de facto background data for use in the statistical procedure

6. When comparing compliance point data to a regulatory standard, no background data is explicitly used. However, in some cases the regulatory standard may be estimated from observed background levels on site

III. STATISTICAL FOUNDATIONS

A. Basic ground-water paradigm:

1. Background and compliance point wells located upgradient and downgradient of potential source of contamination

2. Collect statistically independent samples from compliance wells and background wells on a periodic basis

3. Make statistical comparisons of compliance data to background data, or compliance data to a fixed standard in the permit

4. Must decide each testing period whether contamination has occurred, based on statistical analysis

B. Application of statistical thinking to ground-water setting

1. In ground-water setting, the sample results will vary from period to period even if no release or contamination has occurred. Why?

   a. Variation in lab measurements of concentrations of individual samples

   b. Sampling variability from field collection and handling

c. Natural variation in background levels of pollutants
d. These factors contribute to random variation in sample results that will be observed whether or not contamination has occurred

2. Despite sample fluctuations due to random variation, want to know if the average compliance concentration is significantly higher than the average background concentration

   a. Note that the degree of fluctuation in background and compliance point data relative to the difference in average background and compliance point concentration levels plays a crucial role in distinguishing background behavior from compliance point behavior

   b. Only by careful measurement of sample variability can we accurately make statistical inferences about behavior of overall population (e.g., whether long-term average concentration level at compliance point is greater than background levels)

   c. Because we only get to observe a small sample of the measurement population, forced to use the sample results to describe the overall population characteristics (known as statistical extrapolation or inference)

3. One way to answer above question is to set up hypothesis test using the results of sample ground-water analyses

   a. Hypothesis test makes a decision as to which of two competing notions is closer to the truth, based on the available sample results

   b. Used in groundwater monitoring setting since samples are costly to analyze; only limited data typically available for statistical purposes

   c. Example: Flip a coin 100 times and get all heads

      i. What do we decide about the coin? What is chance of getting heads on next toss?
      ii. Answer: Chance is 100%. Why? Because coin is almost certainly 2-headed!!
      iii. Notion being tested is whether or not coin is fair
      iv. If we say the coin is fair, what evidence can we use to support our claim?
      v. Pr(100 heads in 100 tosses of fair coin)= $(1/2)^{100}$= roughly zero
      vi. But, alternative notion that coin is 2-headed is well supported by evidence at hand
      vii. Key is to determine which hypothesis is best supported by the evidence

4. In ground-water setting, make sure the hypothesis being tested is appropriate to the stage of monitoring or remediation

   a. In detection or compliance monitoring, this becomes H_0: No release versus H_A: Evidence of release, e.g. "innocent until proven guilty" or "clean until proven dirty"

   b. In corrective action, the null hypothesis changes to H_o: "guilty until proven innocent" or "dirty until proven clean"

   c. Choose a statistical test that measures whether the sample data side better with the null hypothesis or the alternative

C. Overview of probability distributions

1. Key to making a hypothesis test work is to describe the mathematical behavior of the sample data. To do this, we often try to fit the sample data to what's known as a probability distribution

2. Because concentrations of pollutants vary in space, over time, etc. in a random but often somewhat regular manner, need to introduce probability models that describe behavior of random variables

   a. Any probability model is only an approximation of the actual physical setting

   b. Often want a model that is mathematically tractable to facilitate statistical calculations

3. Probability Distribution -- Mathematical model to describe the behavior of a random variable

   a. Though we can't predict the next value or measurement, we can attach a probability (i.e., how likely it will occur) to each possible value by using a probability distribution or model

   b. Once a probability model is chosen, we can predict the <u>average</u> behavior of the random variable and other aggregate characteristics

   c. Example: modeling radioactive particles using the Poisson distribution

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

   d. If λ=4, then Pr{X=2}=14.7%. Also, the average value and the standard deviation of X under this model both equal λ or 4. These results follow solely from the mathematical form of the the distribution

D. Normal Distribution

1. Most common example of probability distribution in theory and practice is the familiar bell-shaped curve

2. Normal distribution is a very <u>specific</u> mathematical model

   a. Formula for normal probability model:

$$\Pr\{X = x\} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

   b. Data are called normal only when distributed according to this equation

      i. Under normal curve, about two-thirds of all the measurements will fall within 1 standard deviation (SD) of the average
      ii. Approximately 95% of the measurements will fall within 2 SDs of the average
      iii. Only 5% of the values will fall in the extreme tails beyond 2 SDs

   c. The name "normal" does not imply that other types of data are "abnormal"; lots of typical data do not follow the normal probability model

3. Normal distribution so common because of central limit theorem (CLT)

   a. Theorem states that sums and averages of random variables tend to be normally distributed, even if original random variables <u>do not</u> follow a normal model

   b. Example: Consider distributions of <u>weights</u> of set of people

      i. Each body part has a random weight, all of which must be summed together to find total body weight

ii. Should expect such total weights to be approximately normal

c. Example: Spring-loaded scale leads to model of normal measurement error that is also based on CLT reasoning

d. Because we often use the <u>average</u> of individual sample measurements to estimate the true population average, the normal distribution is quite useful in many statistical tests

i. Normal distribution is also quite mathematically tractable and so has been studied and used extensively in part for this reason

E. Exploratory Data Analysis (EDA)

1. First step in any statistical analysis should be to explore the data for important statistical features and to establish good potential probability models for fitting the measurements

a. Though quite useful, the Normal distribution is not the only candidate model for fitting data; to get correct analysis results, it can be especially important to distinguish when in fact the Normal distribution should not be used

2. EDA consists of basic tools and techniques one can use to decide on good model choices

a. Generally, one wants to perform these calculations via computer software such as GRITS/STAT, GEO-EAS, or MINITAB, but can also explore small data sets by hand computation

F. Initial Graphical Analysis: Time Plots

1. Best initial summary of data is through a graph; picture is worth a thousand data points

2. Time plots offer graphical method to:

a. See all or most of the data simultaneously on one graph

i. Time plots also display the variability in concentration levels over time and can be used to indicate possible outliers

ii. More than one well can be compared on same plot to look for differences between wells

b. View changes in data at a particular well or wells over time

i. Data can easily be examined for trends, perhaps due to changes in background water quality, contamination, etc.

3. How to construct

a. Order well measurements by date of collection

b. Plot analyte measurements on y-axis by date of collection on x-axis

c. GRITS/STAT can construct time plots for one or more wells automatically

G. Central tendency and dispersion

1. Two basic characteristics needed to fit any probability distribution to sample data

a. Average behavior (e.g., mean or median)

b. Sample variability of random variable, i.e. how much it varies from measurement to measurement (e.g., standard deviation (SD) or interquartile range (IQR))

2. Purpose: want to describe typical long-run behavior of random variable

a. Numerical estimates of central tendency:

i. Sample mean: sum of all data divided by number of data points
ii. Sample median: middle point of data after ordering from low to high
iii. Both statistics estimate the typical or average behavior of the data set, however, median is much less influenced than mean by extreme or outlier measurements
iv. Median also less affected by real changes in concentration over time (i.e., contamination)
v. If sample mean and median are very different, data may be skewed or contain notable outliers (sometimes because of contamination)

b. Numerical estimates of variability:

i. Sample variance and standard deviation (SD): variance represents the average squared deviation from the sample mean; standard deviation is the square root of the variance and can be looked at as the typical distance that an individual measurement might be away from the sample mean
ii. Interquartile Range (IQR): difference between the 75th and 25th percentiles of the data; represents the range of values covered by middle 50% of the observed distribution
iii. Both statistics estimate the amount of spread or variability in the data, however, the IQR is much less influenced by outliers
iv. If SD and IQR are quite different, data may be highly skewed or contain significant outliers

H. Graphical Estimates of Central Tendency and Variability: Boxplots

1. Quick way to visualize the distribution of data at one or more wells

a. Basic box plot graphically locates the median, 25th, and 75th percentiles of the data set; some box plots also show other percentiles of the data and/or the minimum and maximum measurements

i. In GRITS/STAT, not only the median but also the sample mean is shown on any box plot
ii. GRITS/STAT version also graphs the minimum and maximum values of each data set

2. Range between the ends of a box plot represents the Interquartile Range (IQR), which can be used as a quick estimate of spread or variability

3. When comparing multiple wells or well groups, can line up box plots for each well side by side on same axes to roughly compare the average and variability in each well

a. Use this technique as a quick exploratory screening for the test of homogeneity of variance across multiple wells, before doing a formal test such as Levene's

b. If two or more boxes are very different in length, the variances in those well groups may be significantly different

c. Box plots may be constructed on original data or residuals of original data

I. Symmetry versus skewness

1. Many types of data are symmetric about a specific value and hence are modeled by symmetric probability distributions like the normal curve

2. Just because distribution is symmetric does not imply normality, however

a. Example: Students t-distribution is symmetric but not normal

b. Reason: t-distribution has fatter tails, hence a greater portion of possible random values are farther from the mean than with normal curve

c. Remember: normal curve is one model with a specific mathematical formula governing the chance/probability that a specific value will be observed. Other models exist that better "fit" certain types of data, and some data don't readily follow any standard probability distribution model

3. Another property important for selecting models for groundwater monitoring data is skewness

a. Normal distribution is symmetric about its mean and median; any negative or positive number is a possible value of a normal random variable

b. Some data distributions are not symmetric, but skewed to one side or the other

c. A skewed distribution is lopsided, with uneven or unbalanced tails

4. Lognormal is great example of distribution with positive skewness and one that serves as a useful model in practice for groundwater monitoring data

a. Income distributions often modelled by lognormal model; most incomes fall in the low to middle class range, with a small percentage of wealthy incomes skewing the right tail

b. In our context, most ground-water data is strictly positive-valued, putting lower bound on possible data range, but leaving upper end open (e.g., minimum concentration=0, maximum concentration=pure product)

c. Most environmental concentration data are found to be approximately lognormal (this includes water quality data)

d. Lognormal distribution gets its name because the logarithms of random lognormal values are normally distributed

i. Key point for later: because of this fact, can often use statistical methods designed for normal data on the logs of lognormal data

5. Numerical estimates of skewness

a. Look at skewness coefficient: represents the average cubed deviation from the sample mean

b. For symmetric distributions, skewness coefficient will be close to zero; for asymmetric distributions, skewness will be either negative or positive depending on whether the left-hand or right-hand tail is longer than its opposite

c. Highly skewed data indicate non-normality; lognormal distribution has positive skewness, normal distribution has zero skewness

d. Coefficient of variation (CV=SD divided by sample mean) sometimes used to indicate skewness, but not as direct or reliable a measure as the skewness coefficient

i. CV of logged measurements particularly unreliable as a measure of skewness, especially when negative values are possible

J. Graphical Assessment of Skewness and Distributional Fit: Probability Plots

1. Very useful visual technique for comparing data to a potential probability model, especially in tails of a distribution

a. Example: often useful for deciding between normal and lognormal models

2. Basic idea: plot ordered sample values versus corresponding expected quantiles or z-scores from N(0,1)

a. Specifically, plot $x_{(i)}$ versus $\Phi^{-1}[i/(n+1)]$

b. Many computer packages such as GRITS/STAT and GEO-EAS will construct p-plots automatically

3. Probability plots (p-plots) can directly show different types of departures from normality

a. Skewness

i. Boxplots can also give graphical indication of skewness

b. Short versus heavy tails

c. Outliers

4. Straight line fit, particularly in the tails, is sign of approximately normal data

a. Compare benzene example data on original versus logged scales

b. Linear fit of logged concentrations suggests original data approximately follow a lognormal distribution

5. Can construct a p-plot with as few as 2 or 3 samples, though such a plot would not be very meaningful.

a. No test of normality will give meaningful results in that case

b. With fewer than 6-8 samples, supplement p-plot with a numerical test of normality such as the Shapiro-Wilk test described later

c. With very small samples, use logs of data under the default assumption that the data are lognormal or substitute a non-parametric technique

IV. BASICS OF HYPOTHESIS TESTING

A. Concept of statistical hypothesis test

1. One basic way to identify statistical evidence of contamination is via an hypothesis test between two competing alternative scenarios:

   a. H_0: No contamination at compliance point

   b. H_A: Contamination has occurred

2. Idea is to decide which alternative is better supported by sample evidence, then make decision based on strength of the data

   a. In ground-water monitoring framework, decide if the compliance point data belong to same population as the background measurements or from a different, more contaminated population

   b. Note: In classical framework, the two competing hypotheses are not given equal weight at the outset

      i. Think of a legal trial to judge someone accused of a crime
      ii. In that case, the competing hypotheses are:

         H_0: Suspect innocent
         H_A: Suspect guilty

      iii. The two choices are not given equal presumption. In fact, we assume H_0 unless the evidence proves "beyond reasonable doubt" that the suspect is guilty
      iv. Much the same is true in statistical hypothesis testing where we demand strong evidence to decide against the null hypothesis H_0

   c. Note: At some RCRA facilities, the usual hypotheses are reversed

      i. Contamination is assumed and the facility must prove that its cleanup procedures have succeeded
      ii. May need a statistical test to measure the changing trend in contamination levels over time

3. Steps involved in hypothesis testing include:

   a. The hypothesis to be tested (often called the null hypothesis H_0) is set up so that—

      i. the observed data or statistic (e.g., the sample mean) will follow a known probability distribution
      ii. the hypothesis represents the assumed condition (e.g., the average concentration of a suspected pollutant of ground water might be assumed to be zero)

   b. An alternative hypothesis (sometimes called the motivating hypothesis, H_A) is set up so that—
      i. under this alternative hypothesis, the data or statistic will follow a distribution different from that under H_0 in a mathematically predictable way
      ii. For example, H_A might specify that the average concentration of a suspected pollutant in ground water is large, leading to a measured concentration greater than zero

c. The measurements are taken and the statistic(s) calculated from the data (e.g., sample mean)

d. The results are compared with the distribution predicted under the null hypothesis H_0. If the probability of the observed result is very small (typically less than 5% or 1%) it must follow that either:

   i. an unlikely event occurred because of random variation (i.e., by chance), or
   ii. the null hypothesis H_0 is incorrect

e. We conclude that H_0 is incorrect, but there is still a possibility that our conclusion is incorrect instead (i.e., H_0 is actually true, but we observed an unlikely event by chance)

4. In-class construction of hypothesis test: Gambler's ruin

5. Since hypothesis test involves deciding between two alternatives, as in a criminal trial, we can make two distinct errors:

   a. Hang the innocent by mistake or free the guilty

   b. In statistical terms, can either:

      i. Accept H_A when H_0 true (false positive)
      ii. Accept H_0 when H_A true (false negative)

   c. The following table lays out the possibilities:

		DECISION	
		Accept H_0	Accept H_A
TRUTH	H_0	OK	False Positive
	H_A	False Negative	OK

6. The two types of error are respectively called Type I and Type II errors

   a. Type I errors known as false positives while Type II errors known as false negatives

   b. Probability of a Type I error is the chance of rejecting a true null hypothesis, denoted by the Greek symbol α

c. In ground-water setting, false positives occur when the statistical test falsely indicates that contamination is present when it is not

d. We can calculate α or set up the hypothesis test to make α a specific value. Alpha is also known as the significance level of the test

e. The probability of a Type II error is the chance of failing to reject a false null hypothesis, denoted by the Greek symbol β

f. False negatives represent cases where the statistical test wrongly fails to identify contamination when it is present

g. Often we work with the complement of β, that is, $1-\beta$, and denote this probability as the power of the test

   i. Power represents the probability that the statistical test will correctly identify contamination when it is present
   ii. Power of a test depends on the significance level α, the amount of data available, and how much the alternative differs from the null hypothesis
   iii. In ground-water setting, statistical power thus depends on how high the average concentration of the pollutant is relative to background; large concentration differences are easier to detect than small ones

7. Illustration: Swedish parking ticket example

a. Person accused of overtime parking because two tires were marked by an officer (say at 2 o'clock and 6 o'clock positions) and found in same position one hour later

b. Defendant claimed that he had moved the car, returned, found same parking space, and that the position of the wheels matched up by pure chance

c. Hypotheses — H_0: Defendant is innocent H_A: Defendant guilty of overtime parking

d. Type I error: chance that judge rules against defendant even though not guilty
Type II error: chance that judge lets defendant off hook even though guilty
Power: chance that judge correctly finds defendant guilty (i.e., that car was not moved as claimed)

e. If each of the 12 hourly positions is equally likely under H_0, and the two wheels rotate independently, the probability of finding the tires in the same positions as the defendant claimed is $(1/12)(1/12)=1/144$ or about 0.0069. This is the Type I error.

   i. Why? Because if the judge finds the defendant guilty, this is the chance that he was actually telling the truth instead

f. Judge ruled that this possibility was likely enough to acquit the defendant (that is, the judge used a significance level for the test of about 0.001, compared to the more usual levels of .05 or .01). Judge also ruled that if all 4 tires had been marked and found later to be in the same position, the evidence would have been strong enough to convict the defendant

B. Nuts and Bolts of Hypothesis Testing: Sampling distributions

1. Key step in making an hypothesis test work is to calculate the chance that the statistic (computed from the data) could have been observed under the null hypothesis, H_0

a. To do this, need to know the distributional behavior of the statistic in question, that is, what would we expect to observe, mathematically, if H_0 is true

2. Idea of sampling distribution:

a. Statistics like the mean or total sum are built from individual random variables and also random in nature, that is, are also random variables

b. By Central Limit Theorem (CLT), when we create new random variables by forming the sum or average of other random variables, the new quantity tends to have an approximately normal distribution, no matter what distribution we started with

c. If we use an average or sum as our test statistic (or sometimes even a function of these quantities), we can expect the test statistic to behave approximately like a normal random variable

i. Warning: CLT applies strictly only to large samples and only approximately to smaller samples; it may not apply to small samples from highly skewed underlying populations

3. But not just any normal distribution

a. Let N(•) denote the normal distribution, μ denote the mean, and σ^2 denote the variance of the original population (that is, suppose the individual measurements are random values from this probability model)

b. Variance of the sampling distribution of the mean depends on the sample size and is not equal to σ^2

c. Example: If we start with $N(\mu,\sigma^2)$, then distribution of the mean of n random observations is $N(\mu, \sigma^2/n)$

i. Note that the variance is reduced by a factor of n
ii. Allows us to predict the behavior of the sample mean with much greater accuracy than the behavior of any single random measurement

4. So why is the sampling distribution of the mean important?

a. Trying to determine characteristics of underlying population based on the limited information contained in the set of collected samples (often to estimate the true population average)

b. Variability of sample mean is much less than that of any single observation, allowing us to better pinpoint the true population mean from the sample average than from any single measurement

c. Example: Suppose we're trying to decide on most appropriate normal model for logged concentrations of benzene at a particular well. Want to know if true mean is μ_1 or μ_2

i. In this example, suppose that μ_1 has been determined from background data levels and the average concentration of compliance point samples at neighboring wells leads to the competing alternative μ_2
ii. If X_1 represents a single observation, it may be hard to decide on basis of X_1 alone which model is best. However, the sampling distribution of the mean tells a much

clearer story, so that if we observe a sample mean of $\bar{x}$ as in picture, we are much more willing to believe that the true average is μ_2 rather than μ_1

iii. Why? Though dispersion of X_1 is large enough to support both models, the variance of $\bar{x}$ is small enough to rule out one model easily

5. Summary

a. Combine individual measurements into an appropriate summary statistic, T_n

b. Sampling distribution describes the statistical behavior of T_n

i. Under H_0, one can gauge whether the calculated value of the statistic is too extreme or unlikely (i.e., is the chance too small?)

ii. Behavior of the statistic under H_0 is critical to performing an hypothesis test

iii. If T_n is too extreme or unlikely, H_0 will be rejected

c. Sometimes T_n does not seem like a natural data summary, but is used because its sampling distribution is known

C. Type I errors: False Positives or False Alarms

1. Definition: Accept H_A when H_0 is true

2. Why should we minimize this?

a. Think of a smoke detector. Only want smoke detector to go off if in fact there is a fire. If the detector is in the kitchen and goes off every time you put something on the stove, you would get a bunch of <u>false alarms</u>

b. Since we want the alarm to go off only when there's an actual fire, we might try to minimize the rate of false alarms or Type I errors by moving detector to another room of the house

3. Now suppose we are running a t-test to decide if average concentration at a particular well is 1 ppb or 5 ppb

a. Set up hypothesis as

H_0: $\mu = 1$ ppb
H_A: $\mu = 5$ ppb

b. Consider sampling distribution of mean concentration, $\bar{x}$. If $\bar{x}$ exceeds a certain critical value, we will decide to reject H_0, that is, if $\bar{x}$ is <u>too big</u>. In other words, reject H_0 if the observed value of $\bar{x}$ is too unlikely under H_0

4. In this setting we choose a decision criterion, the critical point, based on minimizing the Type I error

a. No matter what critical point is chosen, have a small chance of observing a sample mean more extreme than this value coming from the null distribution with true mean μ=1 ppb

b. However, this chance gets smaller and smaller as the critical point is increased

5. Usual strategy is to set α to some small level, say 5% or 1%, and *then choose the critical value* based on the specified α

   a. Why? Though EPA often more concerned with false negatives than false positives, we rely on fixing α because we cannot usually define a minimum magnitude of environmental concern (e.g., the minimum difference in concentration that we want the test to almost certainly detect)

      i. Remember that power, the complement of the false negative rate, depends significantly on the magnitude of the true difference between the null and alternative hypothesis means

   b. If a minimum magnitude of concern existed, a test could be designed to minimize the false negative rate (and maximize the power) instead of the false positive rate

D. Sensitivity (Power) and False Negatives (Type II Errors)

1. Type II error is the chance of a false negative

   a. Example: This sort of error of particular concern in AIDS blood testing

   b. If blood test doesn't detect AIDS antibodies, then AIDS tainted blood might be used on patients

   c. Sometimes have to minimize false negatives more than false positives to be safe

   d. In ground-water testing, false negative means that contamination is present but the test did not identify the ground water as contaminated

2. What happens to Type II error when we choose a critical point?

   a. Since critical point of test is generally chosen by fixing α, false negative rate (β) will depend on two factors:

      i. Difference between the alternative mean and the null mean, relative to the standard error; farther apart the alternative means, smaller the β (since it is easier to detect large differences than small differences)
      ii. Magnitude of the standard error of the sample mean. This in turn depends mostly on sample size; larger the sample size, smaller the β
      iii. Since α is fixed, raising the sample size is generally only way to limit both types of errors simultaneously

3. In ground-water testing, statistical power is the complement of Type II error and equals the probability that the statistical test will find real contamination

   a. Minimizing the false negative rate thus raises the power of the test to reject H_0

   b. Example: Leaky tank study – test specified to minimize Type II error

4. Want statistical power to be as high as possible. Why?

   a. Example: Airport Security Device — imagine the scanner used to detect weapons at airports; device designed to buzz when a metal object is detected

      i. Such devices can be fine-tuned to increase or decrease their sensitivity to metal

b. Naturally want the device to be powerful (i.e., sensitive) enough to pick up substantial metal objects when they exist, for security reasons

c. Don't want device to buzz too infrequently, or weapons might pass through the security checkpoint (i.e., false negative). In this case, one can avoid false negatives (i.e., Type II errors) by increasing the sensitivity of the device to metal objects

d. At same time, don't want the thing to buzz too often or be so sensitive that any minute piece of metal (e.g., non-weapon) will be picked up. That is, don't want too many false positives

5. Airport example leads to general relationships between power, Type I, and Type II errors

a. As sensitivity or power increases, false negatives drop; as sensitivity or power decreases, false negatives increase

b. As sensitivity increases, false alarms tend to increase too, though the relationship isn't precisely linear

c. When applied to ground-water monitoring setting:

i. Type I error — chance that test indicates contamination when it's not there
ii. Type II errors — chance that test fails to indicate real contamination
iii. Statistical Power — chance that test will correctly identify cases of ground-water contamination

E. P-values

1. P-values often used by statisticians as an alternative to critical points for evaluating the results of a statistical test

a. As noted before, one way to set up an hypothesis test is to fix the significance level α (at say 1% or 5%) and compute the critical point necessary to achieve this level

b. When the test statistic exceeds this critical point, the null hypothesis H_0 is rejected and the test fails at the α level of significance

i. Note that the smaller the α, the more extreme the critical point relative to the null distribution mean and the stronger the evidence against H_0

c. Since α represents the chance, precisely at the critical point, that one would observe a test statistic at least that far away from the null distribution mean, the key issue is not how far away the test statistic is but rather what the chance is of observing such an extreme value

2. A p-value represents the probability of observing, under the null hypothesis, a test statistic as or more extreme than that found from the data

a. Rather than compute a critical point, one merely computes the p-value

b. If the p-value is small enough (maybe less than 5%), reject H_0

3. Advantage of a p-value is that the exact chance (e.g., 2.3%) of the result is reported rather than just a given significance (α) level

a. Can compare the relative strength of the results of two different tests more easily using p-values

b. However, difference between using p-values and critical points is really a matter of perspective and not one of significant substance

V. CHECKING ASSUMPTIONS

A. Overview: Statistical models of actual data are approximations of reality. To be used and interpreted properly, the models and statistical test procedures make critical assumptions about the data

1. Why? Many procedures do not perform accurately or efficiently when data don't follow a particular distribution (e.g. normal or lognormal), when outliers are present, when the data are not independent, or when seasonal patterns exist in the data (a type of non-independence)

2. Thus, it is important to check the data to see whether they meet the required assumptions before using a specific statistical test

B. Checking assumptions about the distributional model

1. Cannot confirm that a particular model perfectly fits the data, only that alternative models fit the data more poorly, i.e., model fitting is mostly a process of eliminating bad models. Best we will do is to say a given model provides an approximate fit

2. In ground-water monitoring, checking model assumptions is very important

a. Choice of an appropriate probability model for data can affect the results of statistical tests

b. Example: normal versus lognormal data: If data are lognormal and highly skewed, chances are good that tests on the original data will not show a difference in means even if it exists

i. Use of original data implicitly assumes a normal distribution
ii. Original data too skewed for normal-based tests to pick up differences in population means

c. In such a case, if a significant difference exists, it will only be found by testing the logs of the original data (and implicitly assuming an underlying lognormal population)

3. Guidelines for checking distributional assumptions in ground-water monitoring data analysis

a. If testing for normality of original data, run tests on the raw measurements; if testing for lognormality, run tests of normality on logged groundwater monitoring data

b. Tests for normality should usually be run separately on background well and compliance well data, since the two well populations may have different distributions in the presence of contamination

i. Unless the test is run on the residual values from each well after subtracting the well mean

C. Tests for normality

1. Coefficient of variation (CV) = $s/\bar{x}$ versus skewness coefficient (γ_1)

   a. CV listed in original RCRA guidance document because it is easy to calculate and can be used on small sample sizes, but often not a reliable indication of model appropriateness. Why?

      i. Cutoff in old guidance is to reject normality if the CV is larger than 1
      ii. But even when true coefficient of variation is between (0.5,1), will often get sample CV greater than 1
      iii. Also normal data can sometimes have CV greater than 1, especially if negative data values are possible
      iv. Recommended cutoff of 1 designed to limit the fraction of negative values associated with a probability model of concentration data
      v. When testing for lognormality using the logged measurements, however, may have many negative values. In this case, the CV test can be very misleading

   b. For positive data, CV does give an indication of data skewness, but better to compute sample skewness directly

      i. For normal data, expect a skewness coefficient of zero. Non-normal data will have a positive or negative skewness depending on the type of distribution
      ii. Robustness of t-statistic deteriorates rapidly for γ_1 greater than 1 [Reference: Gayen (1949)]
      iii. Sample skewness γ_1 can be computed approximately as ratio of the average cubed residual to the cube of the standard deviation (SD)
      iv. Many statistical packages will compute the skewness automatically (e.g., GRITS/STAT, GEO-EAS, Minitab)

2. Chi-square test (χ^2)

   a. Also listed in original Interim Final Guidance, but not currently recommended for testing normality. Why?

   b. Most parametric tests like t-test or ANOVA tend to be fairly robust (i.e., valid and efficient) even when the normal assumption fails over the middle ranges of the distribution

   c. Problems occur when the data significantly depart from a normal model in the tails of distribution (e.g., large degree of skewness)

   d. Chi-square test involves dividing sample data into bins/cells based on distinct value ranges and then determining the expected number of observations that should be in each bin assuming a normal distribution

   e. If χ^2 test says data are not normal, it doesn't tell us how they are non-normal, since departures in the middle bins are given same weight or importance as departures from bins representing the tails of distribution

   f. As such, the χ^2 is not as likely to indicate whether some other test procedure for handling non-normal data is really necessary (Reference: Miller, Beyond ANOVA)

   g. Even if there are departures in the tails but the middle part of distribution is fairly normal, χ^2 may not register as significant, whereas some other tests of normality would

3. Probability plots (p-plots) and Filliben's probability plot correlation coefficient

   a. As discussed, a straight line fit on a probability plot is an excellent visual and qualitative indication of normally distributed data

      i. A p-plot cannot be used directly as a statistical test of whether the normal distribution provides an adequate fit to the data; must supplement with a numerical test

   b. One excellent numerical test is the p-plot correlation coefficient

      i. Essentially, a standard correlation is computed between the ordered data and the ordered normal quantiles on the p-plot

   c. If the correlation coefficient is too low, the hypothesis of normality is rejected

      i. Why? A high correlation indicates more of a straight line fit on the p-plot, suggesting that the data are closer to approximate normality, whereas a low correlation indicates a less than straight line fit
      ii. GRITS/STAT automatically provides critical values for the correlation coefficient test at the 5% and 1% significance levels, depending on the sample size

   d. Note: regardless of the fit of the data to the normal distribution, the calculated correlation coefficient is likely to be fairly high by most standards. This happens not because the data fit the model, but because the two sets of values are already ordered from smallest to largest

4. Shapiro-Wilk test of normality (less than 50 data points)

   a. Considered one of the best numerical tests of normality (See Miller, Beyond ANOVA)

      i. Very similar in performance to Filliben's p-plot correlation coefficient
      ii. Can also be used in conjunction with a probability plot to measure how well the plotted quantiles are following a straight line (i.e., how well the sample values are correlated with normal quantiles)

   b. Unlike the Chi-square test, Shapiro-Wilk is most powerful for detecting departures from normality in the tails of a sample distribution

   c. Shapiro-Wilk can be performed on any sample size from 3 to 50 (of course, the power of the test increases as sample size gets larger)

   d. To run:

      i. Order the sample data
      ii. Compute a weighted sum of the differences between the most extreme observations
      iii. Divide the weighted sum by a multiple of the SD and square the result to get Shapiro-Wilk statistic W

   e. Remember that probability plots are a useful supplement to any numerical test of normality; use both the p-plot and a numerical test

      i. Can better visualize data with probability plot
      ii. Can see what type(s) of departure is evident (e.g., outliers, skewness, etc.)

5. Shapiro-Francia test of normality (over 50 data points)

a. Slight modification of Shapiro-Wilk test when sample size is over 50

b. Has same advantages as the Shapiro-Wilk test

c. To run:

i. Order the sample data
ii. Compute a weighted sum of the observations
iii. Divide the square of the weighted sum by a multiple of the SD to obtain the Shapiro-Francia statistic W′

d. The normality of the data is rejected if W′ is too low when compared to the tabulated critical value

D. Overall framework for choosing tests based on distributional assumptions

1. If data are approximately normal or lognormal, use a parametric procedure to analyze the sample data

a. Parametric tests will be the most statistically powerful for detecting concentration differences when the data actually follow the normal or lognormal model

2. If sample data are grossly non-normal and non-lognormal, have one of two options:

a. Find another transformation that leads to normality or another distribution that adequately fits the data

b. Use a non-parametric test based on ranks instead

i. Often must use a rank test when the fraction of nondetects is substantial, because one has difficulty verifying the assumptions of normality or lognormality or finding an adequate alternative transformation or distribution

3. Transformations to normality

a. Have already discussed case of lognormal data, where taking logs of the data gives an approximately normal set of transformed measurements

b. Other transformations can be appropriate for some cases, including square root, reciprocal, cube root, etc.

i. Consultation with a statistician may be required to correctly interpret the results of statistical tests run on data using one of these other transformations

c. Transformations are often performed to get approximate normality and to stabilize the variance of ANOVA residuals across different groups

i. Want to avoid heteroscedasticity or "unequal variances," because approximately equal variances are required for a parametric ANOVA to give valid results

d. Each time a new transformation is tried, the rescaled data can be graphed on a probability plot and tested for normality via one of the tests described above

4. Other distributional models

a. Though normal and lognormal models are the most commonly used, some data may be better fit with an alternative distribution

b. Possible alternate models include the gamma, weibull, and beta distributions

c. Consultation with a statistician will usually be necessary to correctly apply these models in a statistical testing framework

5. Nonparametric rank tests

a. Nonparametric tests don't require any specific distribution for the data and are usually easier to compute

b. However, these tests are less powerful than their parametric counterparts when the data really follow a specific and known probability model

c. However, nonparametric tests are often more powerful than usual tests when data come from an unknown distribution

E. Ensuring that data values are statistically <u>independent</u>

1. Why have independent samples? Because almost all statistical procedures are critically based on the assumption of independence

2. Principally, this is due to the fact that <u>dependent</u> samples (i.e., samples with correlated concentration measurements) will exhibit less variability than really exists in the underlying groundwater population

a. As discussed with sampling distributions, most statistical tests depend on having a good estimate of the true variability in order to make accurate decisions between competing hypotheses

b. Having dependent samples can severely alter the results of hypothesis testing (e.g., consider running CABF t-test with replicate samples)

3. Current guidance recommends that the sampling plan or program at any RCRA facility be developed so that the samples of ground water are physically independent and thus, hopefully, also statistically independent

a. Depending on flow characteristics of site (see Section 3 of Interim Final Guidance), need to allow enough time between samples to ensure that sampling is done on different volumes of ground water

b. Ideally, ANOVA or any statistical procedure that simultaneously tests multiple compliance wells should only be recommended when a site has a higher than average groundwater velocity or where the statistical independence of samples can be guaranteed

c. If physical independence of samples from different wells cannot be assured in quarterly or more frequent sampling episodes, may want to recommend separate interval tests for each compliance well using fewer samples (e.g., collect 1 sample every 6 months)

d. Physical independence <u>does not guarantee</u> statistical independence, although the two will often go hand in hand

4. Testing for statistical independence

a. Basic idea: dataset is probably not statistically independent if the measurements of samples taken closest together in time are strongly correlated (often called serial correlation)

b. One way to estimate the degree of serial correlation in a series of historically collected measurements is to calculate the autocorrelations between neighboring samples

   i. Standard correlation is calculated between the measurement pairs of two variables; autocorrelation is computed on only one variable
   ii. The "measurement pairs" used in computing an autocorrelation consist of samples taken close together in time
   iii. Example: A lag 1 autocorrelation would pair each measurement with the measurement from the very next sampling date and compute a standard correlation on the set of possible pairs of this type (i.e., separated by one sampling date)
   iv. Example: a lag 2 autocorrelation would pair each measurement not with the next sampling date but with the second most recent sampling date; again a standard correlation would be computed on the set of pairs separated by two sampling dates

c. If the serial correlations at all possible lags are zero, the data can be treated as if they are statistically independent

   i. Since the estimated serial correlations are likely to be non-zero even if the true correlations are zero, it can be tricky to decide when the sample serial correlations are small enough to ignore
   ii. Testing formally for non-zero autocorrelation usually requires the key assumption that the data are normally distributed in a specialized way

d. A simpler test that illustrates the concept of independence is the runs count test

   i. Basic idea: in a series of measurements, the individual values should fluctuate around the median in an unpredictable way. Too many consecutive values above or below the overall median (i.e., a "run") is indicative of statistical dependence
   ii. To calculate the runs test, compute the overall sample median, then write down next to a time-ordered list of the data a 1 or 0 for each value, depending on whether the measurement is above or below the median
   iii. Examine the list of 1's and 0's and compute the number of consecutive "runs"
   iv. Example: the list (001110110000) has a total of five runs
   v. If the number of runs is too large (i.e., the data fluctuate up and down in a systematic and non-independent pattern) or too small (i.e., the data exhibit very long runs above or below the median), the hypothesis of independence is rejected

e. Though good conceptually, the runs count test is not as statistically powerful for finding statistical dependence as the rank von Neumann ratio, even though both tests are based on a similar idea

f. Rank von Neumann ratio

   i. Nonparametric test of independence based on the ranks of the data
   ii. To calculate, first rank the data, then list the ranks in the order in which the data were collected
   iii. Compute the von Neumann ratio using the ranks r_i with the following formula:

$$v = \frac{\sum_{i=2}^{n}\left(r_i - r_{i-1}\right)^2}{n\left(n^2 - 1\right)/12}$$

iv. Depending on the sample size, look up the critical points for the von Neumann ratio from one of two tables (listed at the back of this outline)

v. If the computed ratio is either too small or too large, the hypothesis of independence is rejected

F. Correcting for non-independent data in special cases

1. Replicate samples

a. Field replicates or lab splits are not statistically independent measurements and should not be treated as such in statistical procedures

b. Replicate samples tend to be strongly correlated; using replicates as independent data will tend to underestimate overall variability in the ground water population

c. Replicates can be used to measure the component of analytical or sampling variability, just remember that this type of variability is only <u>one</u> of many sources that must be looked at

2. Serial or temporal correlation

a. Serial correlation is present if there exist non-zero autocorrelations in the data; unfortunately, autocorrelation can be generated by many different mathematical processes and models so that it can be very difficult to adequately account for serially dependent data in statistical tests

i. In some cases, one can see simple seasonal patterns in the data (for example, when charting data over time in a time plot)

ii. Cases with simple seasonal fluctuations can be approximately corrected if enough historical data is available (see below)

b. If the degree of serial correlation is strong enough, special allowances must be made in statistical tests, because the estimated variability will be too small and lead to misleading results

i. Example: serial correlation was a major problem with old replicate t-test procedure

ii. Another way to frame the problem: seasonal correlation tends to mask additional variability or noise in the data that is not accounted for by the usual tests, especially if each test includes data from only a limited time period

iii. Why? A limited period of data collection is not enough to sample the full range of ground water population concentrations when the data are serially correlated

c. Correction not needed if the seasonal cycle is long enough

i. If comparisons between background and compliance wells are made every few months, but seasonal patterns fluctuate on the order of a several-year cycle, the effects of serial correlation will tend to be minimal

ii. In that case, point-in-time comparisons of background to compliance data are what really matter in attempting to detect contamination, since the average background concentration level is remaining fairly stable over a period of years

d. Correction when seasonal cycle is less than several years: de-seasonalize data

i. Note that a "season" in statistical terms need not correspond to a usual 3-month season; could be as short as a week or month or much longer

ii. Need to first determine the approximate length of the full seasonal cycle from time plots of the historical data (e.g., 6 months, 1 year, etc.)

e. How to adjust data for a fixed length seasonal cycle

i. Calculate overall mean of historical data set
ii. Compute the seasonal means of all measurements separated by a time lag equal to the full seasonal cycle
iii. Adjust each individual data point by first subtracting the seasonal mean for that sampling date and then adding the overall mean for the whole dataset
iv. Use the adjusted data and not the original data in all subsequent statistical tests
v. Example: data with a yearly cycle

VI. METHODS FOR TWO-SAMPLE COMPARISONS

A. Overview: Two-sample comparisons are appropriate when either a limited number of observations are available at a few wells or there are only two wells to compare

1. Typically, the background or upgradient well data are pooled into one group or sample. Compliance data from one other well makes up the other group.

2. The null hypothesis is that the mean concentration of the pollutant in water samples from the upgradient wells is the same as the mean concentration in the downgradient wells

B. Parametric t-tests

1. The assumptions for a standard, garden variety t-test are:

a. The observations are independent

b. The variances are the same in each group

c. The residuals of each group are normally distributed

d. H_0: the two means are equal

2. To run:

a. Compute residuals in each group by subtracting group mean from each measurement

b. Test residuals for normality and equal variance

c. Compute mean and SD of each group's original measurements

d. If variances are equal, compute t statistic as $t = (\text{Mean}_{\text{down}} - \text{Mean}_{\text{up}})/\text{SE}_{\text{diff}}$

i. SE_{diff} represents the standard error of the difference in sample means
ii. Can be computed using the formula:

$$SE_{diff}=\sqrt{\left[\frac{(n_{up}-1)SD^2_{up}+(n_{down}-1)SD^2_{down}}{n_{up}+n_{down}-2}\right]\left(\frac{1}{n_{up}}+\frac{1}{n_{down}}\right)}$$

e. Compare calculated t statistic with tabulated one-sided critical point $t_c = t_{df,\alpha}$ where df = $(n_{up}+n_{down}-2)$ and α = 1% or 5%

3. If the variances are not equal, the CABF (Cochran's approximation to the Behren's Fisher distribution) t-test procedure may be used

   a. Since the CABF can also be used when the variances are equal and this procedure is built into GRITS/STAT, usually do not need to formally test for equal variances

      i. Still need to test residuals for normality

   b. Key difference in CABF t-test is that the degrees of freedom (df) term is no longer (n_1+n_2-2) but rather a complicated, weighted function of the estimated SDs

4. If the data residuals are not normal but lognormal instead, compute the standard or CABF t-test on the logged data values

   a. Understand what is being tested, however

   b. t-test on logged data is implicitly testing for difference in lognormal medians, not lognormal means

   c. Difference in medians often implies that the means are different too, especially when the variances in the two groups are about equal, but not always

      i. If variances are very different, and CABF t-test is used on logged data, the original means may or may not be significantly different even if medians are different

5. Approximate power of t-test may be found easily for three differences in the two population means ($Mean_{down} - Mean_{up}$):

   a. At a difference of zero, the power is α, the significance level

   b. When the difference in means equals the critical value from the t-table times the standard error of the difference (t_c x SE_{diff}), the power is approximately 50%

   c. When the difference in mean concentrations equals 2 (t_c x SE_{diff}), the power is approximately 100(1-α)%

   d. Can use these rough power calculations to determine how well a particular t-test, given sample size n and significance level α, will be able to find the true mean concentration differences in the previous cases

      i. May need to adjust one or both parameters if more power is needed to detect smaller mean differences

6. If there are many non-detects in the data, use the distribution-free Wilcoxon procedure instead

7. The primary reasons to prefer the standard or CABF t-test to the Wilcoxon are familiarity and the fact that the CABF procedure is still specified in some permits

a. Note that the CABF is a valid procedure in the event that the assumptions are satisfied

b. The problems with the CABF that led to revision of the regulations were

i. It was used when assumptions were substantially violated
ii. It was used with non-independent observations (e.g., multiple aliquots of the same water sample)
iii. It forced the use of a two sample comparison when other procedures were more appropriate to compare more than two groups.

C. Wilcoxon Rank-Sum test for two groups

1. Advantage: test based on ranks of data rather than actual concentrations, hence robust against nonnormality of original values

a. In particular, can be used in presence of a large fraction of nondetects

b. In fact, we recommend use of the Wilcoxon test instead of sign test (recommended in the Interim Final Guidance) even when proportion of non-detects is above 50%

c. Also known as two-sample Mann-Whitney U-Test

2. Basic Algorithm:

a. Assume data are divided into two well groups (e.g., m background samples versus n compliance well samples) with M=m+n total samples

b. Rank the entire set of ordered values lowest to highest as 1 to M; then sum the ranks of samples in the compliance well group and subtract n(n+1)/2 to get Wilcoxon rank-sum statistic W

c. If W is larger than an appropriate critical value, have significant evidence of contamination by Wilcoxon Rank-Sum test

3. Notes on computation and minimum sample sizes

a. Recommended that each well group have at least 4 samples; otherwise Wilcoxon rank-sum test is likely to have very poor power for detecting concentration differences

b. For ease of computation, use a normal approximation to the Wilcoxon statistic with a continuity correction

i. By the CLT, the W statistic has approximately a normal distribution and so we can approximate the exact sampling distribution by a normal density
ii. Continuity correction allows for a better approximation by the continuous normal density of the discrete distribution of rank sums

c. Adjustments for ties (e.g., nondetects)

i. Number of ties at each distinct value must be counted and each tied observation given the same average rank
ii. Approximate variance of Wilcoxon statistic must be adjusted for the tied ranks (see p. 48 of Addendum to Interim Final Guidance)

4. In-Class Wilcoxon Rank-Sum demonstration

a. *Divide class into pairs of arbitrary groups*

b. Have each member count number of coins in their pockets or purses and tabulate results to calculate ranks for each group member

c. Have group members compute basic Wilcoxon statistic and interpret results

D. Why Use Wilcoxon Rank-Sum instead of Sign Test or Test of Proportions?

1. Background of sign test

a. Simple-to-use test for comparing two groups

b. Instead of ranking observations in order, each value treated as 0 or 1 depending on whether it lies below or above median of combined dataset

c. If the proportion of 1's among compliance samples is sufficiently high, can conclude the median concentration of the compliance well is significantly higher than median background concentration

d. Test of proportions is similar in that all compliance samples labeled as 1's and all background samples labeled as 0's

2. Wilcoxon test more powerful than sign test and usually more powerful than test of proportions

a. Though sign test and test of proportions are easy to use, they do not adequately account for differences in concentration magnitudes

b. Since Wilcoxon assigns higher ranks to larger data values, it usually has more statistical power to detect differences between compliance and background levels when they exist

c. When the proportion of non-detects is quite high (>70%), the Wilcoxon test loses its edge in statistical power over the test of proportions. However, the two procedures almost always lead to the same conclusion in those cases, so there is no practical need for the test of proportions (or the sign test)

VII. ANALYSIS OF VARIANCE (ANOVA)

A. Basic Purpose: allow simultaneous comparison of multiple well groups

1. Tests for differences in "average" concentrations levels among all pairs of wells

2. Adjusts for the number of comparisons so that overall false positive rate is kept to a reasonable minimum

a. Use ANOVA instead of running a series of t-tests

B. Parametric One-Way Analysis of Variance (ANOVA)

1. Basic use in groundwater setting is to compare multiple compliance wells against the distribution of background concentrations

2. Main purpose is to assess whether the average concentration at any compliance well is significantly higher than the mean background level

3. Set-up and assumptions

   a. Data must be classified into at least three groups (each group typically a well or group of wells) where the background data comprise the first group

      i. Note: If there are only two groups (i.e., one background group and one downgradient well), use the two-sample t-test
      ii. Should have a bare minimum of 3 to 4 samples per group; much better to have at least 6 to 8 samples per group

   b. Standard parametric ANOVA assumes that the "residuals" are normally distributed and have equal variances across well groups

   c. If either assumption is significantly violated, try one of two options:

      i. If residuals are lognormal instead or the equal variance assumption does not hold, run ANOVA on the logged data
      ii. In this case, the procedure tests for a difference in <u>medians</u> of the original data rather than <u>means</u> of the original data
      iii. However, if variances on logged scale are approximately equal, a difference in medians will also imply a difference in means
      iv. If the logged data also fail the ANOVA assumptions, either try another transformation of the data or use a non-parametric ANOVA, such as the Kruskal-Wallis test described below
      v. Non-parametric ANOVA may particularly be needed if there are a large fraction of non-detect (censored) values

   d. Note that the minimum sample size recommendations are given so that reasonable estimates of the variance can be generated within each group

C. Basic algorithm

1. As long as the fraction of NDs≤15%, set each non-detect to half the detection limit

2. Compute average concentration within each group

3. Compute residual values for each group by subtracting off group mean from each measurement and test assumptions on the residuals

   a. Check normality of all the residuals taken as a whole

   b. Check equality of variances of residuals across well groups

4. If both assumptions concerning the residuals are satisfied, compute the appropriate sum of squares, mean squares, and the F statistic

5. If the F statistic indicates a significant difference among the group means, perform individual comparisons of background data to each compliance well to find the culprit compliance well(s)

6. If either the assumption of normality of residuals or the assumption of equal variances fails, start over with log-transformed data and repeat the above steps

7. If the key assumptions are not met on either the original or logged data (e.g., more than 15% nondetects), perform a non-parametric ANOVA instead

D. How to check assumptions on the residuals

1. Testing normality of residuals

a. Pool all residuals from every well group together

b. Use probability plot on pooled residuals, supplemented by Shapiro-Wilk test or Filliben's probability plot correlation coefficient test

2. Senseless to test normality on original data values instead of the residuals

a. Why? Because if well groups have truly different means (e.g., due to contamination in one or more downgradient wells), overall data may not appear normal though the data may be normal within each separate well group

b. Residuals can be tested because the mean has been removed from each data group (putting each group of data on "equal footing" with mean=0)

3. Testing for equal variance among well groups

a. Homogeneity (equality) of variances of residuals across wells is the most important assumption in parametric ANOVA

i. More important than that of normality of the residuals
ii. Can sometimes still run parametric ANOVA if the equal variance assumption holds but the test of normality barely fails

b. Why? If this assumption is not met, the power of the F-test, that is, its ability to detect differences among the group means, is reduced

i. Mild differences in variances are not too critical
ii. The effect becomes noticeable when the largest and smallest group variances differ by a ratio of about 4; the effect becomes quite severe when the ratio exceeds 10

c. Can use side-by-side box plots to check for equal variances among well groups

i. Quick way to visualize the "spread" or dispersion of the data within a data set

d. To use box plots:

i. Draw box plot of residuals from ANOVA within each group
ii. If box lengths for each group are approximately equal, assume equal variances
iii. If the longest and shortest box lengths differ by ratio of more than 3, use Levene's test of homogeneity to test for significantly different group variances

e. Levene's test of homogeneity of group variances

i. This test is more formal than the box plot approach
ii. This test is not as sensitive to departures from normality as Bartlett's test (discussed in Interim Final Guidance)

g. To run Levene's test

   i. Compute the absolute values of the residuals from each data group
   ii. Compute the F-statistic for an ANOVA on the absolute residuals
   iii. If the calculated F value is less than the tabulated F value, conclude that the variances among the groups are approximately equal
   iv. If the calculated F value exceeds the tabulated F value, conclude that the variances among the groups are not equal, violating the key assumption
   v. If the calculated F fails Levene's test on both the original and logged data, consider running a nonparametric ANOVA

4. What if one or both assumptions fail?

   a. Can try an alternate transformation of the original data values

   b. Use the nonparametric approach by running the Kruskal-Wallis test, which does not require normality of the residuals

E. Interpretation of results of parametric one-way ANOVA

1. If assumptions on residuals check out and we calculate the F statistic along with its significance probability, how do we interpret the results?

2. If the F statistic is not significant, conclude there is no significant difference between the average background level and the average levels of any of the compliance well groups

3. If the F statistic is significant, conclude that at least one pair of well group means is probably different

   a. First do common sense check: look at side-by-side box plots to see if test result seems OK

   b. Note: significant F test does not guarantee that any given pair of means will be significantly different

      i. Enough small but non-significant differences can trigger the cumulative F-test, even though no individual difference is large

   c. Even if the F-test does indicate a difference between two groups, such a result does not guarantee that any single *compliance* well mean is greater than the *background* level

      i. If the average level at any compliance well is less than the background average, the F test may find a difference between two compliance well means but no significant difference between either compliance well and background

   d. Must make individual comparisons between the background data and each individual compliance well to determine which well(s) show evidence of contamination

4. Note on multiple pairwise comparisons

   a. Bonferroni approach when number of comparisons is small (< 6)

      i. Divide the significance level (α) by the number of comparisons

ii. Do a Bonferroni t-test for each comparison at the new α level (i.e., a regular two-sample t-test at the new significance level)

b. When number of comparisons is larger, do t-tests at α=1% level for each comparison to comply with EPA regulation that individual comparisons must be at a significance level of at least 1%

c. For large number of comparisons, might want to abandon ANOVA in favor of a retesting strategy with tolerance or prediction intervals

i. Since the F-statistic is based on cumulative sums, too many non-significant differences can mask one or two significant group differences
ii. With ANOVA applied to a larger monitoring network, could have a single contaminated well missed by the initial F-test
iii. This can happen even with 5 or 10 compliance wells in certain cases

F. Alternative type of parametric ANOVA: Dunnett's multiple comparison with control (MCC) test

1. Method designed to allow comparisons between a single data group (e.g., background) and each of a number of other data groups (e.g., compliance wells)

a. Instead of an overall F test, individual t-tests of each compliance well compared against background are made with special critical points which depend on sample size and the number of compliance wells

b. When assumptions of Dunnett's test are satisfied, the procedure will work better than the usual one-way ANOVA in finding significant differences from background when they exist (i.e., the test will have greater power)

2. Assumptions of Dunnett's MCC test

a. Residuals are normally distributed

b. Equal variances across wells

c. Equal sample sizes in all groups, including background data set

i. Tables of the specialized critical points for the test only exist for the case where each sample size is equal
ii. Can interpolate the approximate critical points using a special scheme when the background sample size is larger than any single compliance well, while each compliance well has the same number of measurements
iii. Need at least 3 observations per well group

3. Basic algorithm of Dunnett's MCC test

a. Letting i index each compliance well and 0 index the background data set, compute t-statistics for each compliance well of the form:

$$t_i = \frac{\sqrt{n}(\bar{y}_i - \bar{y}_0)}{s\sqrt{2}}$$

where n=common group sample size, y_i is the mean of the ith compliance well, y_0 is the mean of the background data set, and s^2 represents the common variance pooled across all (k+1) groups, given by the equation

$$s^2 = \sum_{i=0}^{k} s_i^2 \Big/ (k+1)$$

b. Compare each of the k t-statistics (one for each compliance well) against the Dunnett's test critical point given in table at the end of this outline for α=0.05, and with degrees of freedom equal to k (the number of compliance wells) and υ=(k+1)(n–1)

i. The overall Type I error rate of this procedure is 5%

c. Each t-statistic greater than the critical point represents a compliance well with an average level significantly greater than the mean background level

G. Non-Parametric ANOVA: Kruskal-Wallis test

1. Use to compare several groups of data on a non-parametric basis

a. When comparing only two groups, use the Wilcoxon Rank-Sum test

b. Useful when an ANOVA procedure is desired but the data grossly violate the assumption of normality or when the usual parametric assumptions cannot be easily tested

i. Parametric assumptions can be very difficult to verify in the presence of many non-detects
ii. Recommend switch to Kruskal-Wallis test from the standard ANOVA whenever the fraction of non-detects in the data overall exceeds 15%
iii. Note that the Kruskal-Wallis procedure still assumes that the variances across groups are approximately equal

c. Kruskal-Wallis offers a procedure based on ranked observations that does not depend on the parametric assumption of normality or lognormality of the residuals

2. Basic algorithm

a. Compute the ranks of the combined dataset over all well groups

b. Compute the sum of ranks and the average rank within each group

c. Calculate the Kruskal-Wallis statistic, H, which involves the sum of squared rank-sums and the sample sizes for each well group (see Addendum, p. 44)

i. The statistic H has an approximate chi-square distribution under the null hypothesis of no difference between the average levels of any of the data groups

d. If the Kruskal-Wallis statistic is less than the appropriate chi-square critical value, conclude there are no differences between the median background level and the median levels of the compliance wells

e. If the Kruskal-Wallis statistic is greater than the appropriate chi-square critical value, conclude there are significant differences between the median concentration levels of at least two of the well groups

i. If the Kruskal-Wallis statistic is significant, do individual comparisons between background data and each compliance well group

3. Special Considerations

a. Guidance on sample size and test construction

i. For Kruskal-Wallis test to be sufficiently sensitive to real differences between well groups, it is recommended that the sample size for any group be at least 4
ii. Calculate the degrees of freedom as df= (#groups-1)
iii. Compute the significance probability of the K-W statistic using the table of chi-square critical points in Table 1 on p. B-4 of Interim Final Guidance

b. Presence of tied observations (e.g., nondetects)

i. Compute the number of tied values in each distinct group of ties
ii. Calculate the adjusted Kruskal-Wallis statistic as given on p. 43 of the Addendum to Interim Final Guidance

c. What about the assumption of equal variances?

i. Quick and reasonable check: side-by-side boxplots of <u>ranked</u> measurements
ii. If lengths of boxplots of the ranks are not too different (say less than a ratio of 4 between the longest and shortest length), this assumption should be adequately satisfied

VIII. CONTROL CHARTS

A. Why Think About Control Charts?

1. Control Chart is an alternate method for doing either:

a. Intrawell comparisons, or

b. Comparison of compliance wells to historically-monitored background wells

2. Unlike prediction intervals to be discussed later, Control Charts allow more than point-in-time comparisons of recent data to past information

a. All sample data is continually plotted on a Control Chart as it is collected, providing an historical overview of the concentration pattern at the well and enabling one to see trends or sudden changes in concentration levels over time and to detect possible outliers

b. Control Charts are easy to construct, will show seasonality in the data if present, and can be updated periodically

3. When possible, intrawell comparisons provide the advantage of eliminating worries about spatial variability between wells in different locations

a. Always run the risk when comparing background data to compliance well samples that a significant difference is due to spatial differences between wells at the site rather than actual contamination

b. Intrawell comparisons involve a single well, so that changes in concentration level cannot be spuriously attributed to spatial factors

4. Warning: any intrawell comparison, whether a Control Chart or prediction limit, should only be constructed on initially uncontaminated well measurements

   a. The Control Chart in GRITS/STAT is specifically designed to look for evidence of significant measurement level increases over a baseline measurement

   b. One would not expect to find such an increase if the baseline consists of contaminated ground water samples

   c. Modified versions of the Control Chart or prediction limit procedures can be constructed that specifically attempt to monitor decreases in contamination levels, such as would be needed in a corrective action setting

B. How a Control Chart Works

1. Initial sample information is collected to establish baseline parameters for the Control Chart, specifically, estimates of the well mean and well standard deviation

   a. To gather enough initial information, it is recommended that at least 8 independent samples be collected from prior monitoring before constructing the Control Chart

      i. If an intrawell comparison is being made, prior monitoring must be done at that well
      ii. If comparison to background is being made, prior monitoring would include previous samples from the background wells

   b. These 8 or more samples are not plotted, but only used to estimate the baseline parameters

2. If the baseline data are not independent, but exhibit a seasonal pattern, first deseasonalize the data and use the adjusted data in constructing the Control Chart

3. All future sample data are standardized prior to plotting on the Control Chart, using the baseline parameters

   a. At each sampling period, a standardized mean is computed using the formula

$$Z_i = \frac{\sqrt{n}\left(\bar{x}_i - \mu\right)}{\sigma}$$

   where the formula allows for the collection of more than one sample per sampling period. Alternatively, one can compute the standardized Z_i's after each single sample is collected. In that case, n=1 in the above formula and the mean is replaced by x_i.

   b. Each Z_i is then plotted versus time on the Control Chart

   c. Also a cumulative sum or CUSUM is calculated and plotted on the Control Chart

      i. Compute $S_i = \max\{0, (Z_i-k)+S_{i-1}\}$ where $S_0=0$ is the starting value, i indexes the ith sampling period, and k is a pre-chosen Control Chart parameter
      ii. Calculate and plot one CUSUM for each sampling period
      iii. Note that each CUSUM depends on the CUSUM computed for the previous sampling period

4. Control Chart is declared out-of-control if the sample data become too large relative to the baseline parameters (i.e., when the standardized or CUSUM values cross one of two predetermined threshold values on the Chart)

   a. Idea is that if contamination occurs, the true baseline parameters for the well mean and/or standard deviation will increase, either gradually or sharply

   b. If higher data values (due to contamination) are standardized using the original baseline parameters, the standardized data should start to rise, leading to larger values on the Control Chart and eventually a crossing of a threshold

   c. The thresholds are set so that their crossing signifies a statistically significant result

5. In a combined Shewhart-cumulative sum (CUSUM) Control Chart, like that recommended in the Interim Final Guidance, the Chart is declared out-of-control in one of two ways:

   a. When standardized means (Z_i) computed at each sampling period become too large, crossing the Shewhart control limit (SCL)

      i. Crossing the SCL signifies a rapid rise in well concentration among the recent sample data

   b. When the cumulative sum (CUSUM) of the standardized means becomes too large, crossing the "decision internal value" (h)

      i. Crossing threshold h signifies either a sudden rise in concentration levels or a gradual increase
      ii. A gradual increase is indicated particularly if the CUSUM crosses its threshold but the standardized mean Z_i does not
      iii. Several consecutive small increases in Z_i will not trigger the SCL, but might trigger the CUSUM threshold

6. The recommended threshold values of SCL and h were chosen on the basis of research into the behavior of ground-water monitoring data

   a. Goal of research: establish thresholds which maximize the length of time the process stays "in control" when in fact no contamination is present at the well, and which minimize the length of time the chart stays "out of control" under the same hypothesis (similar to minimizing the Type I error)

   b. Since the Chart is constructed using sample statistics, it has a certain probability of error similar to the α significance level from previous tests

C. Assumptions Behind the Control Chart

1. Data generated by the process, when it is "in control", are Normally distributed

   a. At the very least, initial data used to establish baseline parameters should be tested for Normality
   b. If initial data violate Normality assumption, try a log transformation on the data to see if the assumption is better satisfied. If so, construct the Chart using logged data only

2. Sample data used to construct the Control Chart are independently distributed

a. Control Charts are not very robust (i.e., can give misleading results) when the data are not independent

b. Very important to design the sampling plan in such a way as to collect samples from distinct volumes of water, so as to avoid spatial or seasonal correlation as much as possible

c. Can also test the data for independence and/or seasonality

3. Baseline parameters should reflect current background concentrations levels at the well

a. If the Control Chart reflects an "in control" process for a long period of time, the baseline parameters should be updated to include more recent data as background information

i. The original baseline parameters will be estimated using perhaps 8 prior samples collected during the first year of monitoring
ii. Much better estimates of the true well mean and standard deviation can be obtained by using more data at a later time

b. In general, to update background data with more recent samples, one can run a two-sample t-test comparing the old background levels with the concentrations of the proposed update samples

i. If the t-test does not show a significant difference at the 5% significance level, proceed to update the old background data with more recent sample information (combining all the data into one pool)
ii. If the t-test is significant, however, the new data should not be characterized as background unless some geologic factor can be pinpointed explaining why background levels on the site should have naturally changed
iii. If a geologic reason can be found for the change in background levels, re-estimate the Control Chart baseline parameters using only the more recent sample information

IX. INTERVAL ESTIMATION

A. Another way to test for contamination is to estimate background or compliance well concentrations by constructing a statistical interval

1. Goal: Estimate some characteristic of the population (e.g., average concentration, upper 95th percentile of background data, etc.) or predict future sample values at a well

2. Rationale: Want to estimate an interval because a point estimate tells us nothing about the variability of the statistic. Since any statistic is itself a random variable, very important to know how it might fluctuate

a. Example: Lots of difference between 20 ppm ± 10 ppm and 20 ppm ± 2 ppm

3. In-class experiment to demonstrate the basic characteristics of random intervals

a. Have each person or small group toss a coin 50 times, recording the percent of heads in each group of 10 tosses

b. *Have each group construct a random interval based on* $(p^*_{(2)}, p^*_{(4)})$ as an estimate of true percentage of heads (i.e., probability of tossing H from throw to throw)

d. Plot these intervals by hand on overhead to illustrate that while the true percentage of interest stays the same, each sample of coin tosses leads to a slightly different random interval

4. Random intervals constructed by the <u>same algorithm</u> will change from experiment to experiment though the parameter of interest will not

B. Confidence Levels and Coverage Probabilities

1. Definition of confidence level: proportion of time in the <u>long-run</u> that repeated random intervals will cover the desired parameter

a. Cannot guarantee that the interval constructed for any one sample will cover the parameter, nor can we precisely locate the parameter inside the interval even when it is covered

b. Remember we are using sample statistics to discern features of the overall population, so mistakes are bound to happen

2. Consider example of confidence interval for average background concentration of ground-water monitoring data

a. Formula for confidence interval may be written as:

$$\bar{x} \pm t_\alpha \mathrm{SD}/\sqrt{n}$$

i. In this formula, $(1-\alpha)$=confidence level and α=percent of time we are willing to be dead wrong, i.e., how often the interval will <u>miss</u> the parameter

b. Width of interval indicates the amount of potential error or variability associated with sample average

c. Width depends on three factors:

i. Estimated standard deviation of sample data
ii. Level of confidence chosen beforehand
iii. Sample size

d. To reduce width of a random interval, either

i. Increase sample size, or
ii. Lower the acceptable confidence level

C. Assumptions for Parametric Intervals

1. Standard computing formulas based on having normally distributed data

2. If data is lognormal instead, need to compute confidence intervals on the mean in a fundamentally different way than tolerance or prediction intervals

a. In all cases, however, will need to compute sample mean and SD on the logged data values

3. Can also construct a nonparametric interval when assumptions are not met, but these generally require greater amounts of data to construct an interval with equivalent degrees of confidence and/or coverage

D. Computing Parametric Statistical Intervals

1. General formula is of form: $\bar{x} \pm \kappa\sigma$

2. Necessary components include sample mean, sample standard deviation, and the factor κ

3. Factor κ depends on the type of interval being constructed, sample size, and the confidence level desired (see **Section 4** of the Addendum to Interim Final Guidance)

E. Confidence Intervals

1. Overview

a. Most common type puts bounds on the true average concentration in a groundwater population

i. Other types include confidence intervals for upper percentiles of the population

b. Only recommended for two compliance monitoring scenarios:

i. When the fixed limit is an ACL determined from average background concentration data
ii. When the fixed limit is a risk-based MCL or ACL

c. Interpretation of a 95% confidence interval on the population mean: "I'm 95% sure that the true mean concentration is contained between these interval limits"

2. If comparing compliance data to a Ground-Water Protection Standard (GWPS) that has been established on the basis of average background well concentrations

a. Compute confidence interval on mean of compliance point data

b. Compare GWPS to the lower limit of confidence interval

c. If GWPS is below the lower confidence limit, one has evidence of a possible violation

3. If comparing compliance data to a risk-based ACL or MCL:

a. Compute confidence interval on upper 95th percentile of compliance point data

b. Compare ACL/MCL to lower limit of confidence interval

c. If lower confidence limit is above ACL/MCL, one has evidence of a possible contamination

4. Computing Confidence Intervals

a. Simplest case: mean of normal data

i. First compute sample mean and standard deviation

ii. Given confidence level (1–α) and sample size n, let factor κ equal to $t_{\alpha,n-1}/\sqrt{n}$

iii. Lower limit computed as $\bar{x} - \kappa \cdot \mathrm{SD}$

b. Mean of lognormal data

i. Compute sample mean and SD of logged data

ii. Use Land's (1971) formulas to compute upper and lower confidence limits:

$$LL_{\alpha} = \exp\left(\bar{y} + 0.5s_y^2 + \frac{s_y H_{\alpha}}{\sqrt{n-1}}\right) \qquad UL_{1-\alpha} = \exp\left(\bar{y} + 0.5s_y^2 + \frac{s_y H_{1-\alpha}}{\sqrt{n-1}}\right)$$

c. 95th percentile of normal data

i. Compute sample mean and SD

ii. Use table from Hahn and Meeker (1993) to find factor κ (provided at back of outline)

iii. Lower limit computed in this case as $\bar{x} + \kappa \cdot \mathrm{SD}$ (note the plus sign and the fact that the upper 95th percentile will almost always be larger than the mean)

d. 95th percentile of lognormal data

i. Compute same interval as above on logged data, then exponentiate the lower and upper limits to find confidence interval for original data

ii. Lower limit becomes $\exp\left[\bar{y} + \kappa \cdot \mathrm{SD}_y\right]$ where y is used to denote the logged data

5. Minimum sample sizes

a. To construct the interval need enough observations to generate an adequate estimate of the sample variability

b. Recommend at least 4 data points at the very minimum, though the interval is likely to be extremely wide unless closer to 8 to 10 observations are used

6. Note on interpretation

a. A confidence interval on the sample mean only estimates the approximate level of the true concentration average

b. Such an interval does not tell us where the upper 95th percentile of the concentration distribution lies

c. For example, in comparison with a GWPS it may happen that the entire confidence interval for the mean lies below the limit, yet some individual samples have values above the compliance standard

d. Be sure to choose the correct type of confidence interval

F. Tolerance Intervals

1. Overview

   a. Appropriate when using upper percentile of concentration distribution to gauge compliance

   b. One-sided tolerance interval estimates an upper bound on a large fraction of the possible concentration measurements

      i. Definition of coverage: percentage of all population measurements included within the tolerance interval

   c. Note: tolerance intervals will generally be wider than confidence intervals about the mean since the sample mean will have less variability than the distribution as a whole

   d. Used in detection monitoring when tolerance interval is computed on background data and compared to individual compliance point samples

   f. Interpretation of a one-sided 95% tolerance interval with 95% coverage: "I'm 95% sure that approximately 95% of individual population measurements fall below this upper limit

2. If comparing upgradient versus downgradient wells (as in detection monitoring):

   a. Compute tolerance interval on background data

   b. When the sample size is small to moderate, if any single compliance point sample exceeds the upper 95% tolerance limit, one has significant evidence that the background and compliance well concentration distributions are different, indicating evidence of contamination

   c. When the sample size is larger, expect 1 in every 20 samples to fail an upper 95% tolerance limit just by chance

3. Do not use tolerance intervals for compliance monitoring

   a. Recent guidance suggests comparing upper tolerance limit to GWPS

   b. But since the upper tolerance limit is equal to the <u>upper</u> confidence limit on the 95th percentile, this comparison is likely to produce more frequent false positives

   c. Much better to compare <u>lower</u> confidence limit on the 95th percentile to the GWPS

   d. Note: comparison of interest is whether the GWPS is exceeded by more than a specified fraction of the compliance concentrations (e.g., 5% for the case of 95% coverage)

4. Computing tolerance intervals on normal data

   a. First compute sample mean and standard deviation on background samples

   b. Then compute factor κ for a one-sided upper tolerance limit with 95% minimum coverage

      i. Use Table 5 on p. B-9 in Interim Final Guidance or the Hahn/Meeker table listed at back of this outline

   c. Set upper tolerance limit equal to $\bar{x} + \kappa \cdot SD$

5. *Computing tolerance intervals on lognormal data*

   a. First compute sample mean and standard deviation on logged background samples

   b. Then compute factor κ for a one-sided upper tolerance limit with 95% minimum coverage

      i. Use Table 5 on p. B-9 in Interim Final Guidance or the Hahn/Meeker table listed at back of this outline

   c. Set upper tolerance limit equal to $\bar{x} + \kappa \cdot SD$

   d. Exponentiate the logged upper tolerance limit to get a tolerance limit on the original data scale

6. Minimum sample size requirements

   a. Tolerance interval can be computed with as few as 3 data values; however, to have a passable estimate of the standard deviation, one should usually have at least 8-10 background samples

7. Non-parametric tolerance limits

   a. Use when data show evidence of non-normality or high proportion of nondetects

   b. Easy to construct: set upper tolerance limit to maximum of background samples

      i. Can be useful in retesting scenarios

   c. Based on the number of samples available, one can compute the expected minimum or average coverage of the tolerance limit (see **Section 4** of the Addendum)

      i. Because the non-parametric tolerance limit makes fewer assumptions about the data, more samples are typically needed to achieve the same coverage at the 95% confidence level than with parametric tolerance limits

G. Prediction Intervals

1. Overview:

   a. As opposed to confidence and tolerance intervals, which estimate specific characteristics of the overall population (e.g., mean, 95th percentile, 95% coverage), a prediction interval estimates bounds on the concentrations of future samples, specifically the next k future samples

   b. Used in detection monitoring by constructing prediction interval from background data and comparing future compliance observations against the upper limit to see if contamination is indicated

   c. Used in intrawell comparisons by computing interval from past data at the well to predict expected values of k future well samples

   d. The number of future samples, k, may be as small as one

   e. Interpretation of 95% prediction interval: "I'm 95% sure that the next k future sample values will fall below the upper prediction limit"

2. When comparing upgradient versus downgradient well data:

   a. Compute prediction interval on background data

   b. If any one or more of the k compliance samples exceeds the upper prediction limit, one has significant evidence of contamination

3. When making intrawell comparisons:

   a. Compute prediction interval on past well data

   b. If any or more of the k new samples exceeds upper prediction limit, have significant evidence of recent contamination

4. Computing prediction intervals

   a. First calculate sample mean and standard deviation

   b. Determine number of future samples (k) to be collected during next sampling period

   c. Then calculate factor κ as

$$\kappa = t_{n-1,.05/k}\sqrt{1+\frac{1}{n}}$$

   where the t-value is taken from a standard t-distribution

   d. Note: if data are lognormal, construct interval on logged data and then exponentiate the end result to get a prediction limit on the original scale

5. Minimum sample size requirements

   a. Similar to those for a confidence interval, since a reasonable estimate of the standard deviation must be computed from the past data

   b. The number of future samples is arbitrary, but the number of past data must be at least 4 and should be closer to 8 or more

6. Notes on interpretation

   a. Prediction intervals will generally give wider limits than comparable confidence intervals based on same data, but limits that are often shorter in width than a tolerance interval

   b. Advantage: only need at least one compliance point sample to compare to an upper prediction limit, so prediction intervals can be applied to slow moving ground water where the independence of samples separated even by months at a time is difficult to ensure

7. Non-parametric prediction limits

   a. Use when data show evidence of non-normality or high proportion of nondetects

   b. Easy to construct

      i. When comparing background wells to compliance wells, set upper prediction limit to maximum of background samples

ii. When doing intrawell comparisons, set upper prediction limit to maximum of past well data

c. Based on the number of samples available, one can compute the confidence level associated with the upper prediction limit (see **Section 4** of the Addendum)

i. Because the non-parametric prediction limit makes fewer assumptions about the data, more samples are typically needed to achieve the same confidence level than with parametric prediction limits

H. Summary: how do we choose between the interval types?

1. Basic criterion is the existing permit if it provides specific guidance

2. If not, consider the type of data available and type of monitoring being done

a. Remember, the type of interval used can make a huge difference in the resulting decision—in general, the widths of confidence, tolerance, and prediction intervals will be very different on the same sample data

3. Differences between the interval types

a. Statistical intervals have different uses depending on the purpose in mind

b. Hahn's astronaut example (1970): An astronaut awaiting his tour of duty on the space shuttle is not concerned about what happens on average during such flights (confidence interval), nor with what happens on 95% of all flights (tolerance interval), but rather with what will happen on his or her specific flights (prediction interval)

c. Roulette wheel at a casino: A player is concerned with what he or she will win on the next few bets (prediction interval); the casino owners care about their average winnings in order to make a profit (confidence interval); while the wheel operator who makes a commission on each bet lost by a player is concerned about the long-run proportion of lost bets (tolerance interval)

X. NONDETECTS AND OUTLIERS

A. Need strategies for treating nondetect values

1. Nondetects occur frequently with many ground-water monitoring parameters

B. Recommend the following overall decision framework:

1. With 15% or fewer nondetects, make a simple substitution of one-half the detection limit (DL)

a. Note: samples with estimated concentrations below the DL should be treated as valid measurements for statistical purposes and not replaced by one-half the DL

2. With more than 15% nondetects

a. If ANOVA or t-test procedure to be run, switch to non-parametric alternative like the Kruskal-Wallis or Wilcoxon procedures

i. Note when ranking nondetects that "detected but not quantified" samples should be given higher ranks than nondetect samples which are "undetected"

b. If an interval test to be run and the nondetect fraction is less than 50%, try either Cohen's or Aitchison's adjustment in order to estimate a parametric interval

i. If Cohen's and Aitchison's methods fail, try to construct a non-parametric interval

c. If the fraction of nondetects is over 50%, switch to non-parametric test or construct a Poisson-based prediction or tolerance limit

d. If all samples are nondetect, no statistical test needs to be run

C. Adjustments to Parametric Intervals

1. Cohen's adjustment

a. Assumes the observed data (detects and nondetects) come from the same, censored distribution (i.e., nondetects have low, but positive, concentrations)

b. Based on the censored probability model, we estimate a new mean and standard deviation

c. The adjusted mean and standard deviation can be substituted into the formula for a prediction or tolerance limit

2. Aitchison's adjustment

a. Assumes that detects come from one distribution but nondetects represent zero concentrations

b. Like Cohen's method, the assumption of a particular probability model for the data leads to adjusted estimates of the mean and standard deviation

3. Deciding between Cohen's and Aitchison's methods

a. Important to decide on an appropriate model for the dataset

b. Compare Censored pp-plot against Detects-only pp-plot to help

c. Also consider the parameter being monitored and the physical aspects of the RCRA facility

D. Poisson-based Intervals

1. Only consider when the fraction of nondetects is quite high, say 90% or more

2. Poisson-based methods use information about concentration magnitudes of detects even though most samples are nondetect

a. Data values are not ranked; instead the original measurements are used

b. Nondetect samples can be replaced by DL/2

3. Poisson-based prediction limits

   a. Goal is to compute an upper prediction limit that contains the sum of the next k future measurements

   b. If sum of future measurements exceeds the prediction limit, one has evidence of a violation

   c. Calculate the upper Poisson prediction limit as

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

   where k=number of future samples, n=number of background samples, c=k/n, T_k*=sum of k future concentrations, T_n=sum of n background measurements, and z_α is the upper α percentile of the standard Normal distribution

4. Poisson-based tolerance limits (see **Section 2.2.5** of Addendum)

E. Outlier (extreme value) testing

1. Definition: A constituent value that is very different from most other values in the data set for the same ground-water constituent

2. Possible reasons:

   a. Contaminated sampling equipment

   b. Inconsistent sampling or analytical chemistry methodology resulting in laboratory contamination or other anomalies

   c. Errors in the transcription of data values or decimal points

   d. True but extreme measurements

3. Formal testing for outliers should be done only if an observation(s) seems particularly high (by orders of magnitude) compared to the rest of the data set

4. Once an observation is found to be an outlier, the following action should be taken:

   a. If the error can be identified and the correct value can be recovered, replace the outlier value with the corrected value

   b. If the error can be documented but the corrected value cannot be recovered, the outlier should be deleted. Describe this deletion in the statistical report

   c. If no error can be documented, then assume that the value is a valid measurement. Do not remove it from the data set and do not alter it. Try to obtain another sample to confirm the high value

5. Procedure to test for outlier(s)

a. Purpose: To determine whether there is statistical evidence that an observation that appears extreme does not fit the distribution of the rest of the data

b. Assumptions for outlier tests

i. The data set, excluding the suspect data point, come from a normal distribution
ii. Since ground-water data often follow a lognormal distribution, may need to sometimes run outlier test on log-transformed data

c. Before running test, use probability plots excluding the suspected outlier(s) to see whether data seem to be more normal or more lognormal

d. To run:

i. Order the data from smallest to largest so that suspect data is the largest value
ii. Calculate the mean and SD of <u>all</u> the data, including the outlier
iii. Compute T_n, the difference between the largest observation and the sample mean, divided by the SD
iv. Compare the calculated statistic T_n to the tabulated value
v. If the calculated value exceeds the tabulated value, there is evidence that the suspect observation is a statistical outlier

XI. <u>MULTIPLE COMPARISONS AND RETESTING STRATEGIES</u>

A. Power Curves

1. Since the power always equals (1-ß), just as Type II error depends on how far apart the null and alternative means are, so does the power

2. Power of a statistical test for detecting differences tends to increase as the alternative is "farther" from the null hypothesis

3. Most of the time, we are testing hypotheses of form:

a. H_0: $\mu \leq \mu_0$ vs. H_1: $\mu > \mu_0$

b. μ_0 might be the mean background level or a fixed compliance limit in the permit

c. In these cases, want to reject H_0 when the evidence indicates the true compliance mean is something larger than μ_0, regardless of the specific alternative value

d. Hence, we are interested in whole range of alternative hypotheses, collectively called a compound alternative

4. Under a compound alternative hypothesis, very useful method of analyzing the performance of specific statistical test is through examination of its power curve

a. Power curve is graph of the test's power/sensitivity over the range of possible alternatives

5. Power curve allows us to decide between different tests in the range of alternatives of most concern

   a. Generally, the more powerful the test, given the same false alarm rate α, the better the test for statistical purposes

B. Problem of Multiple Comparisons in Ground-water Testing

1. Background:

   a. Most of the comments on testing so far have dealt with a single comparison of one data set (background well) versus another data set (compliance well)

   b. The Type I and II errors are based on running a single statistical test

   c. Situation often different at RCRA facilities with multiple compliance wells and several parameters to test per well

   d. If one test is run for each well and parameter combination, can substantially increase the false positive rate; that is, the rate that the test indicates contamination when in fact no contamination exists

2. To illustrate how this can happen do class experiment

   a. Have each participant roll dice 20 times and count number of sixes

   b. Test H_0: Pr{roll a 6} = 1/6 vs H_A: Pr{roll a 6} > 1/6 by following criterion:

      i. Expect only about 3 sixes in 20 tosses
      ii. Reject H_0 if S = # sixes in 20 tosses is greater than 5

   c. Using this criterion, the false positive rate should only be 10%

   d. Determine the number of class who reject test

   e. Since Pr{roll a six} = 1/6 in reality, any one who rejects H_0 has a false positive on their hands

3. These results should illustrate the more general situation:

   a. Though the error rate of any single test is low, if we run enough tests the chance of making at least one mistake (false positive) overall is high

   b. In general, Pr{at least one false (+)} = $1-(1-\alpha)^{wc}$ where w=# wells and c=# constituents being tested

   c. In fact, for 100 tests run each at 5% Type I error rate:

      i. Expect on average 100 * .05 = 5 false positives
      ii. Pr{at least one false (+)}= $1-(.95)^{100}$ = 99.4%

4. Experimentwise versus Comparisonwise Error Rates

a. Previous example illustrates difference between a comparisonwise error rate, i.e., the false positive rate associated with single well comparison (e.g., 5% in hypothetical example, with $w=1$, $c=1$) and

b. The experimentwise error rate, which represents the overall expected error rate based on the total number of statistical tests being run (total # of tests=$w \cdot c$)

5. Basic problem is that for large numbers of compliance wells and constituents, if statistical tests are run on all cases, the error rate is likely to be very high even when no contamination is occurring

a. Thus, even a small facility is likely to see at least one significant test result during statistical testing, forcing further sampling and lab analysis to verify the result or even forcing the facility into compliance monitoring on basis of false alarm

b. Because of this, want to design our statistical procedures to keep the overall experimentwise error rate down to an acceptable minimum without sacrificing power necessary to detect actual contamination

6. Important reminder: False positive rate only makes sense if in fact no contamination has occurred at any of these wells. Real contamination should never be mistaken for a false alarm

C. Strategies to Handle Multiple Comparisons and Lower Overall Experimentwise Error Rate

1. For relatively small number of individual comparisons (≤ 5), we can lower/adjust the Type I error rate to account for the number of tests run (Bonferroni approach)

2. Rationale:

a. Lowering the α for individual tests reduces the overall experimentwise error to acceptable levels

b. Distributes the probability of occurrence of a false positive evenly among all wells in the experiment

3. Example of Bonferroni approach

a. Suppose we need to run 5 tests and original single comparison α set to 5%

b. Adjust alpha by running each comparison at $\alpha^*=\alpha/5=1\%$

c. Then experimentwise false alarm rate drops from 22% to 5%

4. For larger number of comparisons, Bonferroni approach cannot be used directly

a. EPA regulations mandate that any single comparison have an error rate of at least 1%

b. Why? Because lowering the false alarm rate generally lowers the overall sensitivity and power of test

c. Recall Airport security gate: to lower the false alarm rate, have to make sensor buzz less often, which also lowers detection rate of dangerous objects

d. Not desirable since EPA wants tests to have power to detect actual contamination when it occurs

5. Instead, second basic strategy is to use Omnibus Testing

   a. Data from several wells are grouped together and tested as a whole, at an error rate of 5%, usually using some type of ANOVA (Analysis of Variance) procedure

   b. If overall test shows a significant difference between compliance and background concentrations, individual comparisons are conducted to determine which well or wells from the group are contaminated

   c. Advantage is that fewer statistical tests need to be run when no contamination is actually present

   d. Disadvantage when contamination is present: must do further post-hoc testing to determine which well(s) is contaminated

6. Alternate strategy: retesting

   a. Instead of omnibus tests, an alternative is to retest each well that tests positive for contamination prior to moving into compliance monitoring

   b. California example: recent proposal sets up a testing regimen involving comparison of two independent samples against a prediction interval constructed from background data, both statistical tests run at α=1%

   c. If either retest shows a significant difference, the well is said to be contaminated, but if both retests are not significant, no contamination is inferred (the original sample result is classified as spurious)

   d. Advantage is that the overall false positive rate remains small even when many wells are tested, yet power of test is comparable to usual testing procedures

D. Retesting strategies recommended in Addendum to Interim Final Guidance

1. Rationale for development of other strategies

   a. One characteristic of California retesting proposal (and its relation to the EPA standard prediction interval limit) is that its power curve depends on the number of wells in the downgradient network

   b. Might want greater flexibility in choosing a retesting strategy, perhaps tailored to the type of network being tested

2. Need to meet two basic goals when selecting a particular testing strategy

   a. Keep overall facility-wide false positive rate low (say approximately 5%, regardless of the number of downgradient wells)

   b. Maintain effective power comparable to the EPA Reference Power Curve

      i. Effective power refers to the statistical power of a testing strategy to correctly identify contamination at exactly one and only one well within a network of multiple wells (that is, one well is contaminated but the rest aren't)

      ii. The EPA Reference Power Curve is the power curve associated with using a 99% confidence level upper prediction limit to test the next single future measurement at exactly one downgradient well

iii. Note that the EPA Reference Power Curve does not depend on the number of wells in the overall network, but does depend on the number of background samples used to construct the upper prediction limit

3. Parametric retesting strategies

a. Collect and pool background data from appropriate wells

b. Construct a 95% confidence level upper tolerance limit with given average coverage level (specified below) and an upper prediction limit with given confidence level (also specified below), both based on the same set of background measurements

c. Compare one new sample from each compliance point well to the upper tolerance limit

d. If any well triggers the tolerance limit, collect one or more resamples from that well and compare the resample(s) to the upper prediction limit constructed above

e. Make a decision about each suspect compliance well on the basis of the resamples:

i. If all resamples pass the upper prediction limit, conclude that the original sample was high by random fluctuation and that the well is still clean
ii. If any resample fails the upper prediction limit, conclude that the well shows significant evidence of a higher-than-background average concentration level

f. Key task in picking the right strategy: choose the tolerance limit coverage and the prediction limit confidence level so that the twin goals of minimizing the Type I error rate and maintaining adequate statistical power are met

i. Have to consider the number of wells in the monitoring network and also the number of background samples available
ii. See table on p. 70 of Addendum to Interim Final Guidance for recommended choices under some possible scenarios
iii. Further simulation of the effective power may be necessary to pick a strategy for networks not listed on p. 70 of the Addendum; modify the SAS code listed in Appendix B of the Addendum to tailor the program to specific networks

4. Non-parametric retesting strategies

a. When the background data used to construct parametric tolerance and prediction intervals don't satisfy the usual parametric assumptions (e.g., when there are many nondetects), non-parametric prediction limits may be constructed instead

b. First construct an upper non-parametric prediction limit on the background data (usually the maximum observed measurement)

c. Compare one new sample from each compliance well to the upper prediction limit

d. For each compliance well that triggers the prediction limit, collect one or more resamples from that well depending on the number of background samples available and the number of downgradient wells in the overall network (see below)

e. Compare the resample(s) from each suspect compliance well to the original upper prediction limit and make a decision:

i. If all resamples pass the upper prediction limit, classify the well as clean for that testing period

ii. If any resample fails the upper prediction limit, classify the well as having a significantly elevated concentration level compared to background

f. To decide on an appropriate number of resamples to take (in order to meet the basic goals with regard to Type I error and effective power), see table on p. 74 of Addendum to Interim Final Guidance for several possible scenarios

i. Simulation may again be necessary to decide on an appropriate for networks not listed in this table; modify the SAS code in Appendix B of the Addendum as needed

g. Unlike using parametric intervals, one can only improve the effective power of the non-parametric retesting strategy by either:

i. Increasing the number of background samples used to construct the upper prediction limit, or

ii. Taking more resamples from each suspect compliance well (unfortunately, there are severe practical limits on this option due to the need for statistically independent data)

Seasonal Data

	Number of Data Points n = 48			Number of Data Points n = 48	
Station 1 Year	Month	Station 1	Station 2 Year	Month	Station 2
1	1	6.32	1	1	6.29
1	2	6.08	1	2	6.11
1	3	5.16	1	3	5.66
1	4	4.47	1	4	5.16
1	5	4.13	1	5	4.75
1	6	3.65	1	6	6.79
1	7	3.48	1	7	4.51
1	8	3.78	1	8	4.37
1	9	3.94	1	9	4.95
1	10	4.40	1	10	5.22
1	11	4.94	1	11	5.73
1	12	5.32	1	12	6.72
2	1	5.82	2	1	7.42
2	2	5.76	2	2	7.56
2	3	4.88	2	3	6.13
2	4	4.84	2	4	6.24
2	5	4.87	2	5	5.07
2	6	4.13	2	6	4.95
2	7	3.51	2	7	4.59
2	8	4.32	2	8	5.22
2	9	4.06	2	9	5.13
2	10	4.47	2	10	5.69
2	11	5.05	2	11	6.41
2	12	5.20	2	12	7.53
3	1	5.83	3	1	7.02
3	2	5.65	3	2	6.93
3	3	5.32	3	3	6.55
3	4	5.33	3	4	6.66
3	5	4.20	3	5	6.69
3	6	3.85	3	6	5.23
3	7	4.45	3	7	5.14
3	8	3.56	3	8	5.06
3	9	3.85	3	9	5.71
3	10	4.72	3	10	6.17
3	11	5.38	3	11	6.78
3	12	5.33	3	12	7.64
4	1	6.59	4	1	7.46
4	2	5.93	4	2	7.56
4	3	4.98	4	3	7.30
4	4	4.61	4	4	7.22
4	5	4.18	4	5	6.07
4	6	3.79	4	6	5.53
4	7	3.64	4	7	5.65
4	8	3.77	4	8	5.94
4	9	4.05	4	9	6.68
4	10	4.50	4	10	6.42
4	11	5.15	4	11	7.10
4	12	5.57	4	12	7.86

Critical Values for the Rank von Neumann Ratio Test.

	Alpha				
N	*.005*	*.010*	*.025*	*.050*	*.100*
4	—	—	—	—	.66
5	—	—	—	.70	.88
6	.354	.475	.63	.81	1.02
7	.502	.573	.71	.87	1.11
8	.549	.626	.78	.94	1.15
9	.575	.672	.83	.99	1.20
10	.62	.72	.89	1.04	1.23
11	.67	.77	.93	1.08	1.26
12	.71	.81	.96	1.11	1.29
13	.74	.84	1.00	1.14	1.32
14	.78	.87	1.03	1.17	1.34
15	.81	.90	1.05	1.19	1.36
16	.84	.93	1.08	1.21	1.38
17	.87	.96	1.10	1.24	1.40
18	.89	.98	1.13	1.26	1.41
19	.92	1.01	1.15	1.27	1.43
20	.94	1.03	1.17	1.29	1.44
21	.96	1.05	1.18	1.31	1.45
22	.98	1.07	1.20	1.32	1.46
23	1.00	1.09	1.22	1.33	1.48
24	1.02	1.10	1.23	1.35	1.49
25	1.04	1.12	1.25	1.36	1.50
26	1.05	1.13	1.26	1.37	1.51
27	1.07	1.15	1.27	1.38	1.51
28	1.08	1.16	1.28	1.39	1.52
29	1.10	1.18	1.30	1.40	1.53
30	1.11	1.19	1.31	1.41	1.54
32	1.13	1.21	1.33	1.43	1.55
34	1.16	1.23	1.35	1.45	1.57
36	1.18	1.25	1.36	1.46	1.58
38	1.20	1.27	1.38	1.48	1.59
40	1.22	1.29	1.39	1.49	1.60
42	1.24	1.30	1.41	1.50	1.61
44	1.25	1.32	1.42	1.51	1.62
46	1.27	1.33	1.43	1.52	1.63
48	1.28	1.35	1.45	1.53	1.63
50	1.29	1.36	1.46	1.54	1.64
55	1.33	1.39	1.48	1.56	1.66
60	1.35	1.41	1.50	1.58	1.67
65	1.38	1.43	1.52	1.60	1.68
70	1.40	1.45	1.54	1.61	1.70
75	1.42	1.47	1.55	1.62	1.71
80	1.44	1.49	1.57	1.64	1.71
85	1.45	1.50	1.58	1.65	1.72
90	1.47	1.52	1.59	1.66	1.73
95	1.48	1.53	1.60	1.66	1.74
100	1.49	1.54	1.61	1.67	1.74

One-Sided Critical Points for Dunnett's Test

(One-tailed) $d^{\alpha}_{k,\nu}$

	α = .05									α = .01								
ν \ k	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
5	2.02	2.44	2.68	2.85	2.98	3.08	3.16	3.24	3.30	3.37	3.90	4.21	4.43	4.60	4.73	4.85	4.94	5.03
6	1.94	2.34	2.56	2.71	2.83	2.92	3.00	3.07	3.12	3.14	3.61	3.88	4.07	4.21	4.33	4.43	4.51	4.59
7	1.89	2.27	2.48	2.62	2.73	2.82	2.89	2.95	3.01	3.00	3.42	3.66	3.83	3.96	4.07	4.15	4.23	4.30
8	1.86	2.22	2.42	2.55	2.66	2.74	2.81	2.87	2.92	2.90	3.29	3.51	3.67	3.79	3.88	3.96	4.03	4.09
9	1.83	2.18	2.37	2.50	2.60	2.68	2.75	2.81	2.86	2.82	3.19	3.40	3.55	3.66	3.75	3.82	3.89	3.94
10	1.81	2.15	2.34	2.47	2.56	2.64	2.70	2.76	2.81	2.76	3.11	3.31	3.45	3.56	3.64	3.71	3.78	3.83
11	1.80	2.13	2.31	2.44	2.53	2.60	2.67	2.72	2.77	2.72	3.06	3.25	3.38	3.48	3.56	3.63	3.69	3.74
12	1.78	2.11	2.29	2.41	2.50	2.58	2.64	2.69	2.74	2.68	3.01	3.19	3.32	3.42	3.50	3.56	3.62	3.67
13	1.77	2.09	2.27	2.39	2.48	2.55	2.61	2.66	2.71	2.65	2.97	3.15	3.27	3.37	3.44	3.51	3.56	3.61
14	1.76	2.08	2.25	2.37	2.46	2.53	2.59	2.64	2.69	2.62	2.94	3.11	3.23	3.32	3.40	3.46	3.51	3.56
15	1.75	2.07	2.24	2.36	2.44	2.51	2.57	2.62	2.67	2.60	2.91	3.08	3.20	3.29	3.36	3.42	3.47	3.52
16	1.75	2.06	2.23	2.34	2.43	2.50	2.56	2.61	2.65	2.58	2.88	3.05	3.17	3.26	3.33	3.39	3.44	3.48
17	1.74	2.05	2.22	2.33	2.42	2.49	2.54	2.59	2.64	2.57	2.86	3.03	3.14	3.23	3.30	3.36	3.41	3.45
18	1.73	2.04	2.21	2.32	2.41	2.48	2.53	2.58	2.62	2.55	2.84	3.01	3.12	3.21	3.27	3.33	3.38	3.42
19	1.73	2.03	2.20	2.31	2.40	2.47	2.52	2.57	2.61	2.54	2.83	2.99	3.10	3.18	3.25	3.31	3.36	3.40
20	1.72	2.03	2.19	2.30	2.39	2.46	2.51	2.56	2.60	2.53	2.81	2.97	3.08	3.17	3.23	3.29	3.34	3.38
24	1.71	2.01	2.17	2.28	2.36	2.43	2.48	2.53	2.57	2.49	2.77	2.92	3.03	3.11	3.17	3.22	3.27	3.31
30	1.70	1.99	2.15	2.25	2.33	2.40	2.45	2.50	2.54	2.46	2.72	2.87	2.97	3.05	3.11	3.16	3.21	3.24
40	1.68	1.97	2.13	2.23	2.31	2.37	2.42	2.47	2.51	2.42	2.68	2.82	2.92	2.99	3.05	3.10	3.14	3.18
60	1.67	1.95	2.10	2.21	2.28	2.35	2.39	2.44	2.48	2.39	2.64	2.78	2.87	2.94	3.00	3.04	3.08	3.12
120	1.66	1.93	2.08	2.18	2.26	2.32	2.37	2.41	2.45	2.36	2.60	2.73	2.82	2.89	2.94	2.99	3.03	3.06
∞	1.64	1.92	2.06	2.16	2.23	2.29	2.34	2.38	2.42	2.33	2.56	2.68	2.77	2.84	2.89	2.93	2.97	3.00

Values of $H_\alpha = H_{0.05}$ for Computing a One-Sided Lower 95% Confidence Limit on a Lognormal Mean

	n									
s_y	3	5	7	10	12	15	21	31	51	101
0.10	-1.130	-1.806	-1.731	-1.690	-1.677	-1.666	-1.655	-1.648	-1.644	-1.642
0.20	-1.949	-1.729	-1.678	-1.653	-1.646	-1.640	-1.636	-1.636	-1.637	-1.641
0.30	-1.816	-1.669	-1.639	-1.627	-1.625	-1.625	-1.627	-1.632	-1.638	-1.648
0.40	-1.717	-1.625	-1.611	-1.611	-1.613	-1.617	-1.625	-1.635	-1.647	-1.662
0.50	-1.644	-1.594	-1.594	-1.603	-1.609	-1.618	-1.631	-1.646	-1.663	-1.683
0.60	-1.589	-1.573	-1.584	-1.602	-1.612	-1.625	-1.643	-1.662	-1.685	-1.711
0.70	-1.549	-1.560	-1.582	-1.608	-1.622	-1.638	-1.661	-1.686	-1.713	-1.744
0.80	-1.521	-1.555	-1.586	-1.620	-1.636	-1.656	-1.685	-1.714	-1.747	-1.783
0.90	-1.502	-1.556	-1.595	-1.637	-1.656	-1.680	-1.713	-1.747	-1.785	-1.826
1.00	-1.490	-1.562	-1.610	-1.658	-1.681	-1.707	-1.745	-1.784	-1.827	-1.874
1.25	-1.486	-1.596	-1.662	-1.727	-1.758	-1.793	-1.842	-1.893	-1.949	-2.012
1.50	-1.508	-1.650	-1.733	-1.814	-1.853	-1.896	-1.958	-2.020	-2.091	-2.169
1.75	-1.547	-1.719	-1.819	-1.916	-1.962	-2.015	-2.088	-2.164	-2.247	-2.341
2.00	-1.598	-1.799	-1.917	-2.029	-2.083	-2.144	-2.230	-2.318	-2.416	-2.526
2.50	-1.727	-1.986	-2.138	-2.283	-2.351	-2.430	-2.540	-2.654	-2.780	-2.921
3.00	-1.880	-2.199	-2.384	-2.560	-2.644	-2.740	-2.874	-3.014	-3.169	-3.342
3.50	-2.051	-2.429	-2.647	-2.855	-2.953	-3.067	-3.226	-3.391	-3.574	-3.780
4.00	-2.237	-2.672	-2.922	-3.161	-3.275	-3.406	-3.589	-3.779	-3.990	-4.228
4.50	-2.434	-2.924	-3.206	-3.476	-3.605	-3.753	-3.960	-4.176	-4.416	-4.685
5.00	-2.638	-3.183	-3.497	-3.798	-3.941	-4.107	-4.338	-4.579	-4.847	-5.148
6.00	-3.062	-3.715	-4.092	-4.455	-4.627	-4.827	-5.106	-5.397	-5.721	-6.086
7.00	-3.499	-4.260	-4.699	-5.123	-5.325	-5.559	-5.886	-6.227	-6.608	-7.036
8.00	-3.945	-4.812	-5.315	-5.800	-6.031	-6.300	-6.674	-7.066	-7.502	-7.992
9.00	-4.397	-5.371	-5.936	-6.482	-6.742	-7.045	-7.468	-7.909	-8.401	-8.953
10.00	-4.852	-5.933	-6.560	-7.168	-7.458	-7.794	-8.264	-8.755	-9.302	-9.918

Source: After Land, 1975

Values of $H_\alpha = H_{0.10}$ for Computing a One-Sided Lower 90% Confidence Limit on a Lognormal Mean

	n									
s_y	3	5	7	10	12	15	21	31	51	101
0.10	-1.431	-1.320	-1.296	-1.285	-1.281	-1.279	-1.277	-1.277	-1.278	-1.279
0.20	-1.350	-1.281	-1.268	-1.266	-1.266	-1.266	-1.268	-1.272	-1.275	-1.280
0.30	-1.289	-1.252	-1.250	-1.254	-1.257	-1.260	-1.266	-1.272	-1.280	-1.287
0.40	-1.245	-1.233	-1.239	-1.249	-1.254	-1.261	-1.270	-1.279	-1.289	-1.301
0.50	-1.213	-1.221	-1.234	-1.250	-1.257	-1.266	-1.279	-1.291	-1.304	-1.319
0.60	-1.190	-1.215	-1.235	-1.256	-1.266	-1.277	-1.292	-1.307	-1.324	-1.342
0.70	-1.176	-1.215	-1.241	-1.266	-1.278	-1.292	-1.310	-1.329	-1.349	-1.370
0.80	-1.168	-1.219	-1.251	-1.280	-1.294	-1.311	-1.332	-1.354	-1.377	-1.403
0.90	-1.165	-1.227	-1.264	-1.298	-1.314	-1.333	-1.358	-1.383	-1.409	-1.439
1.00	-1.166	-1.239	-1.281	-1.320	-1.337	-1.358	-1.387	-1.414	-1.445	-1.478
1.25	-1.184	-1.280	-1.334	-1.384	-1.407	-1.434	-1.470	-1.507	-1.547	-1.589
1.50	-1.217	-1.334	-1.400	-1.462	-1.491	-1.523	-1.568	-1.613	-1.063	-1.716
1.75	-1.260	-1.398	-1.477	-1.551	-1.585	-1.624	-1.677	-1.732	-1.790	-1.855
2.00	-1.310	-1.470	-1.562	-1.647	-1.688	-1.733	-1.795	-1.859	-1.928	-2.003
2.50	-1.426	-1.634	-1.751	-1.862	-1.913	-1.971	-2.051	-2.133	-2.223	-2.321
3.00	-1.560	-1.817	-1.960	-2.095	-2.157	-2.229	-2.326	-2.427	-2.536	-2.657
3.50	-1.710	-2.014	-2.183	-2.341	-2.415	-2.499	-2.615	-2.733	-2.864	-3.007
4.00	-1.871	-2.221	-2.415	-2.596	-2.681	-2.778	-2.913	-3.050	-3.200	-3.366
4.50	-2.041	-2.435	-2.653	-2.858	-2.955	-3.064	-3.217	-3.372	-3.542	-3.731
5.00	-2.217	-2.654	-2.897	-3.126	-3.233	-3.356	-3.525	-3.698	-3.889	-4.100
6.00	-2.581	-3.104	-3.396	-3.671	-3.800	-3.949	-4.153	-4.363	-4.594	-4.849
7.00	-2.955	-3.564	-3.904	-4.226	-4.377	-4.549	-4.790	-5.037	-5.307	-5.607
8.00	-3.336	-4.030	-4.418	-4.787	-4.960	-5.159	-5.433	-5.715	-6.026	-6.370
9.00	-3.721	-4.500	-4.937	-5.352	-5.547	-5.771	-6.080	-6.399	-6.748	-7.136
10.00	-4.109	-4.973	-5.459	-5.920	-6.137	-6.386	-6.730	-7.085	-7.474	-7.906

Source: After Land, 1975

Factors for Calculating Normal Distribution One-Sided $100(1-\alpha)\%$ Tolerance Bounds and Confidence Intervals for Percentiles (Part I)

		p = 0.900					p = 0.950					p = 0.990					
n	1 - α:	0.010	0.025	0.050	0.100	0.200	0.010	0.025	0.050	0.100	0.200	0.010	0.025	0.050	0.100	0.200	n
2		-0.707	-0.143	0.138	0.403	0.737	0.000	0.273	0.475	0.717	1.077	0.564	0.761	0.954	1.225	1.672	2
3		-0.072	0.159	0.334	0.535	0.799	0.295	0.478	0.639	0.840	1.126	0.782	0.958	1.130	1.361	1.710	3
4		0.123	0.298	0.444	0.617	0.847	0.443	0.601	0.743	0.922	1.172	0.924	1.088	1.246	1.455	1.760	4
5		0.238	0.389	0.519	0.675	0.883	0.543	0.687	0.818	0.982	1.209	1.027	1.182	1.331	1.525	1.801	5
6		0.319	0.455	0.575	0.719	0.911	0.618	0.752	0.875	1.028	1.238	1.108	1.256	1.396	1.578	1.834	6
7		0.381	0.507	0.619	0.755	0.933	0.678	0.804	0.920	1.065	1.261	1.173	1.315	1.449	1.622	1.862	7
8		0.431	0.550	0.655	0.783	0.952	0.727	0.847	0.958	1.096	1.281	1.227	1.364	1.493	1.658	1.885	8
9		0.472	0.585	0.686	0.808	0.968	0.768	0.884	0.990	1.122	1.298	1.273	1.406	1.530	1.688	1.904	9
10		0.508	0.615	0.712	0.828	0.981	0.804	0.915	1.017	1.144	1.313	1.314	1.442	1.563	1.715	1.922	10
11		0.538	0.642	0.734	0.847	0.933	0.835	0.943	1.041	1.163	1.325	1.349	1.474	1.591	1.738	1.937	11
12		0.565	0.665	0.754	0.863	1.004	0.862	0.967	1.062	1.180	1.337	1.381	1.502	1.616	1.758	1.950	12
13		0.589	0.685	0.772	0.877	1.013	0.887	0.989	1.081	1.196	1.347	1.409	1.528	1.638	1.776	1.962	13
14		0.610	0.704	0.788	0.890	1.022	0.909	1.008	1.098	1.210	1.356	1.434	1.551	1.658	1.793	1.973	14
15		0.629	0.721	0.802	0.901	1.029	0.929	1.026	1.114	1.222	1.364	1.458	1.572	1.677	1.808	1.983	15
16		0.647	0.736	0.815	0.912	1.036	0.948	1.042	1.128	1.234	1.372	1.479	1.591	1.694	1.822	1.992	16
17		0.663	0.750	0.827	0.921	1.043	0.965	1.057	1.141	1.244	1.379	1.499	1.608	1.709	1.834	2.000	17
18		0.678	0.763	0.839	0.930	1.049	0.980	1.071	1.153	1.254	1.385	1.517	1.625	1.724	1.846	2.008	18
19		0.692	0.775	0.849	0.939	1.054	0.995	1.084	1.164	1.263	1.391	1.534	1.640	1.737	1.857	2.015	19
20		0.705	0.786	0.858	0.946	1.059	1.008	1.095	1.175	1.271	1.397	1.550	1.654	1.749	1.867	2.022	20
21		0.716	0.796	0.867	0.953	1.064	1.021	1.107	1.184	1.279	1.402	1.565	1.667	1.761	1.876	2.028	21
22		0.728	0.806	0.876	0.960	1.068	1.033	1.117	1.193	1.286	1.407	1.579	1.680	1.772	1.885	2.034	22
23		0.738	0.815	0.884	0.966	1.073	1.044	1.127	1.202	1.293	1.412	1.592	1.691	1.782	1.893	2.039	23
24		0.748	0.823	0.891	0.972	1.076	1.054	1.136	1.210	1.300	1.416	1.605	1.702	1.791	1.901	2.045	24
25		0.757	0.831	0.898	0.978	1.080	1.064	1.145	1.217	1.306	1.420	1.616	1.713	1.801	1.908	2.049	25
26		0.766	0.839	0.904	0.983	1.084	1.074	1.153	1.225	1.311	1.424	1.628	1.723	1.809	1.915	2.054	26
27		0.774	0.846	0.911	0.988	1.087	1.083	1.161	1.231	1.317	1.427	1.638	1.732	1.817	1.922	2.058	27
28		0.782	0.853	0.917	0.993	1.090	1.091	1.168	1.238	1.322	1.431	1.648	1.741	1.825	1.928	2.063	28
29		0.790	0.860	0.922	0.997	1.093	1.099	1.175	1.244	1.327	1.434	1.658	1.749	1.833	1.934	2.067	29
30		0.797	0.866	0.928	1.002	1.096	1.107	1.182	1.250	1.332	1.437	1.667	1.757	1.840	1.940	2.070	30
35		0.828	0.893	0.951	1.020	1.108	1.141	1.212	1.276	1.352	1.451	1.708	1.793	1.871	1.965	2.087	35
40		0.854	0.916	0.970	1.036	1.119	1.169	1.236	1.297	1.369	1.462	1.741	1.823	1.896	1.986	2.101	40
50		0.894	0.950	1.000	1.059	1.134	1.212	1.274	1.329	1.396	1.480	1.793	1.869	1.936	2.018	2.122	50
60		0.924	0.976	1.022	1.077	1.146	1.245	1.303	1.354	1.415	1.493	1.833	1.903	1.966	2.042	2.138	60
120		1.020	1.059	1.093	1.134	1.184	1.352	1.395	1.433	1.478	1.535	1.963	2.016	2.063	2.119	2.189	120
240		1.092	1.121	1.146	1.175	1.211	1.431	1.463	1.492	1.525	1.565	2.061	2.100	2.135	2.176	2.227	240
480		1.145	1.166	1.184	1.205	1.231	1.491	1.514	1.535	1.558	1.588	2.134	2.163	2.189	2.218	2.255	480
∞		1.282	1.282	1.282	1.282	1.282	1.645	1.645	1.645	1.645	1.645	2.326	2.326	2.326	2.326	2.326	∞

Factors for Calculating Normal Distribution One-Sided 100(1-α)% Tolerance Bounds and Confidence Intervals for Percentiles (Part II)

		p = 0.900					p = 0.950					p = 0.990					
n	1 - α:	0.800	0.900	0.950	0.975	0.990	0.800	0.900	0.950	0.975	0.990	0.800	0.900	0.950	0.975	0.990	n
2		5.049	10.253	20.581	41.201	103.029	6.464	13.090	26.260	52.559	131.426	9.156	18.500	37.094	74.234	185.617	2
3		2.871	4.258	6.155	8.797	13.995	3.604	5.311	7.656	10.927	17.370	5.010	7.340	10.553	15.043	23.896	3
4		2.372	3.188	4.162	5.354	7.380	2.968	3.957	5.144	6.602	9.083	4.110	5.438	7.042	9.018	12.387	4
5		2.145	2.742	3.407	4.166	5.362	2.683	3.400	4.203	5.124	6.578	3.711	4.666	5.741	6.980	8.939	5
6		2.012	2.494	3.006	3.568	4.411	2.517	3.092	3.708	4.385	5.406	3.482	4.243	5.062	5.967	7.335	6
7		1.923	2.333	2.755	3.206	3.859	2.407	2.984	3.399	3.940	4.728	3.331	3.972	4.642	5.361	6.412	7
8		1.859	2.219	2.582	2.960	3.497	2.328	2.754	3.187	3.640	4.285	3.224	3.783	4.354	4.954	5.812	8
9		1.809	2.133	2.454	2.783	3.240	2.268	2.650	3.031	3.424	3.972	3.142	3.641	4.143	4.662	5.389	9
10		1.770	2.066	2.355	2.647	3.048	2.220	2.568	2.911	3.259	3.738	3.078	3.532	3.981	4.440	5.074	10
11		1.738	2.011	2.275	2.540	2.898	2.182	2.503	2.815	3.129	3.556	3.026	3.443	3.852	4.265	4.829	11
12		1.711	1.966	2.210	2.452	2.777	2.149	2.448	2.736	3.023	3.410	2.982	3.371	3.747	4.124	4.633	12
13		1.689	1.928	2.155	2.379	2.677	2.122	2.402	2.671	2.936	3.290	2.946	3.309	3.659	4.006	4.472	13
14		1.669	1.895	2.109	2.317	2.593	2.098	2.363	2.614	2.861	3.189	2.914	3.257	3.585	3.907	4.337	14
15		1.652	1.867	2.068	2.264	2.521	2.078	2.329	2.566	2.797	3.102	2.887	3.212	3.520	3.822	4.222	15
16		1.637	1.842	2.033	2.218	2.459	2.059	2.299	2.524	2.742	3.028	2.863	3.172	3.464	3.749	4.123	16
17		1.623	1.819	2.002	2.177	2.405	2.043	2.272	2.486	2.693	2.963	3.841	3.137	3.414	3.684	4.037	17
18		1.611	1.800	1.974	2.141	2.357	2.029	2.249	2.453	2.650	2.905	2.822	3.105	3.370	3.627	3.960	18
19		1.600	1.782	1.949	2.108	2.314	2.016	2.227	2.423	2.611	2.854	2.804	3.077	3.331	3.575	3.892	19
20		1.590	1.765	1.926	2.079	2.276	2.004	2.208	2.396	2.576	2.808	2.789	3.052	3.295	3.529	3.832	20
21		1.581	1.750	1.905	2.053	2.241	1.993	2.190	2.371	2.544	2.766	2.774	3.028	3.263	3.487	3.777	21
22		1.572	1.737	1.886	2.028	2.209	1.983	2.174	2.349	2.515	2.729	2.761	3.007	3.233	3.449	3.727	22
23		1.564	1.724	1.869	2.006	2.180	1.973	2.159	2.328	2.489	2.694	2.749	2.987	3.206	3.414	3.681	23
24		1.557	1.712	1.853	1.985	2.154	1.965	2.145	2.309	2.465	2.662	2.738	2.969	3.181	3.382	3.640	24
25		1.550	1.702	1.838	1.966	2.129	1.957	2.132	2.292	2.442	2.633	2.727	2.952	3.158	3.353	3.601	25
26		1.544	1.691	1.824	1.949	2.106	1.949	2.120	2.275	2.421	2.606	2.718	2.937	3.136	3.325	3.566	26
27		1.538	1.682	1.811	1.932	2.085	1.943	2.109	2.260	2.402	2.581	2.708	2.922	3.116	3.300	3.533	27
28		1.533	1.673	1.799	1.917	2.065	1.936	2.099	2.246	2.384	2.558	2.700	2.909	3.098	3.276	3.502	28
29		1.528	1.665	1.788	1.903	2.047	1.930	2.089	2.232	2.367	2.536	2.692	2.896	3.080	3.254	3.473	29
30		1.523	1.657	1.777	1.889	2.030	1.924	2.080	2.220	2.351	2.515	2.684	2.884	3.064	3.233	3.447	30
35		1.502	1.624	1.732	1.833	1.957	1.900	2.041	2.167	2.284	2.430	2.652	2.833	2.995	3.145	3.334	35
40		1.486	1.598	1.697	1.789	1.902	1.880	2.010	2.125	2.232	2.364	2.627	2.793	2.941	3.078	3.249	40
50		1.461	1.559	1.646	1.724	1.821	1.852	1.965	2.065	2.156	2.269	2.590	2.735	2.862	2.980	3.125	50
60		1.444	1.532	1.609	1.679	1.764	1.832	1.933	2.022	2.103	2.202	2.564	2.694	2.807	2.911	3.038	60
120		1.393	1.452	1.503	1.549	1.604	1.772	1.841	1.899	1.952	2.015	2.488	2.574	2.649	2.716	2.797	120
240		1.358	1.399	1.434	1.465	1.501	1.733	1.780	1.819	1.854	1.896	2.437	2.497	2.547	2.591	2.645	240
480		1.335	1.363	1.387	1.408	1.433	1.706	1.738	1.766	1.790	1.818	2.403	2.444	2.479	2.509	2.545	480
∞		1.282	1.282	1.282	1.282	1.282	1.645	1.645	1.645	1.645	1.645	2.326	2.326	2.326	2.326	2.326	∞

STATISTICAL ANALYSIS OF GROUND-WATER MONITORING DATA AT RCRA FACILITIES

DRAFT

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

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

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.

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.

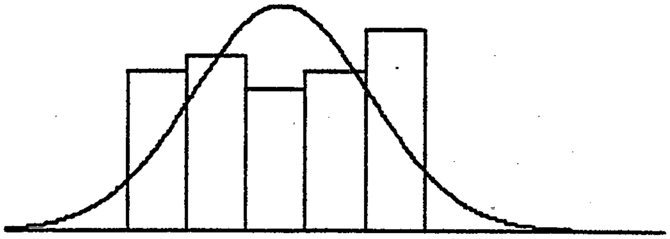
The Coefficient-of-Variation is recommended within the Interim Guidance because it is easy to calculate and is amenable to small sample sizes. To ensure that a Normal model which predicts a significant fraction of negative concentration values is not fitted to positive data, the Interim Final Guidance 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

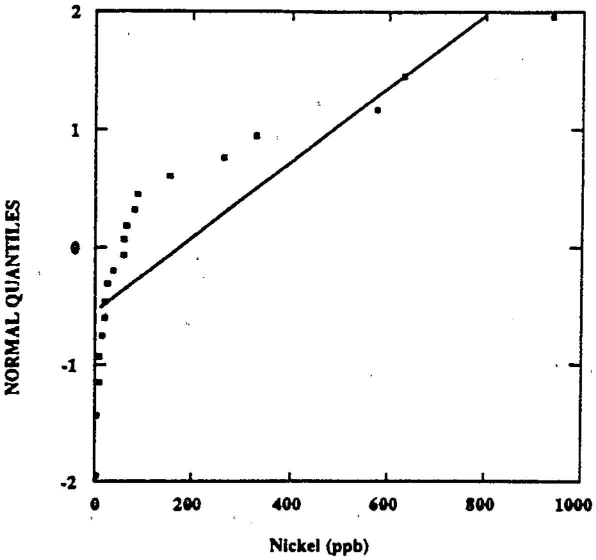
Step 1. List the measured nickel concentrations in order from lowest to highest.

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:

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

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:

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

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

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

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≤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):

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

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

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

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

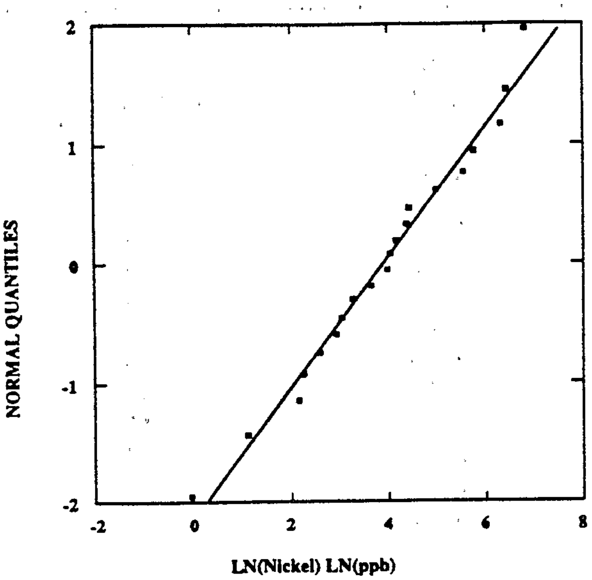
Step 1. List the natural logarithms of the measured nickel concentrations in order from lowest to highest.

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.

PROBABILITY PLOT



Method 2. Coefficient of Skewness

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

$$\overline{x} = 3.918 \log(\text{ppb})$$

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

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

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

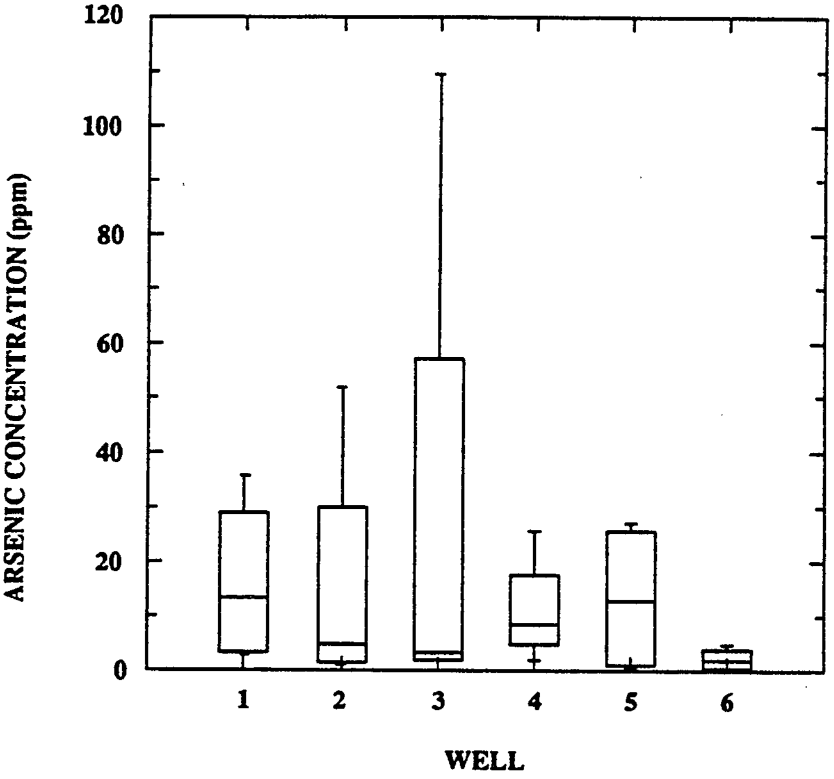
SOLUTION

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

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

where x_{ij} represents the jth value from the ith well and $\overline{x}_i$ 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 $(\overline{x}_i)$

Well 1 mean = 16.47

Well 2 mean = 15.76

Well 3 mean = 29.60

Well 4 mean = 11.26

Well 5 mean = 13.49

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 not recommended 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 × (%NDs) and n × (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.

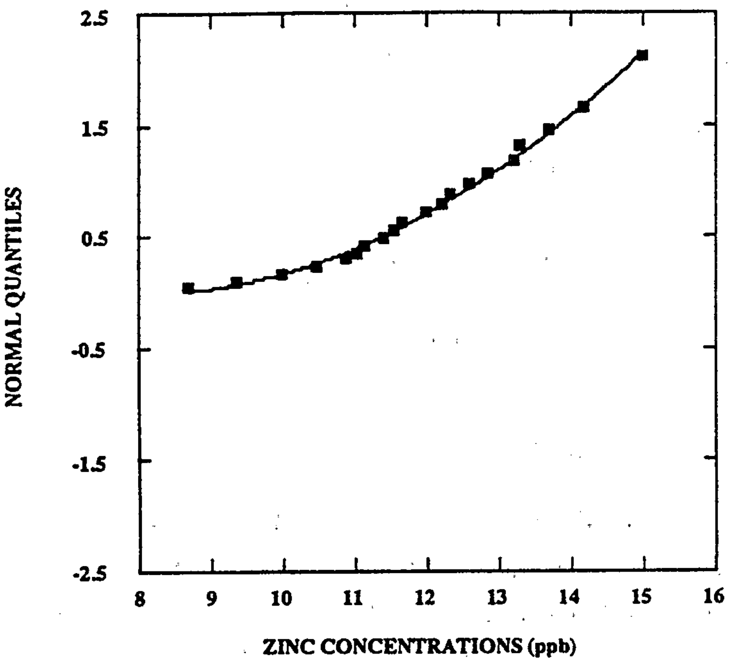
Order (i)	Zinc Conc. (ppb)	Censored Probs.	Normal Quantiles	Detects-Only Probs.	Normal Quantiles
1	<7	.024	-1.971		
2	<7	.049	-1.657		
3	<7	.073	-1.453		
4	<7	.098	-1.296		
5	<7	.122	-1.165		
6	<7	.146	-1.052		
7	<7	.171	-0.951		
8	<7	.195	-0.859		
9	<7	.220	-0.774		
10	<7	.244	-0.694		
11	<7	.268	-0.618		
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

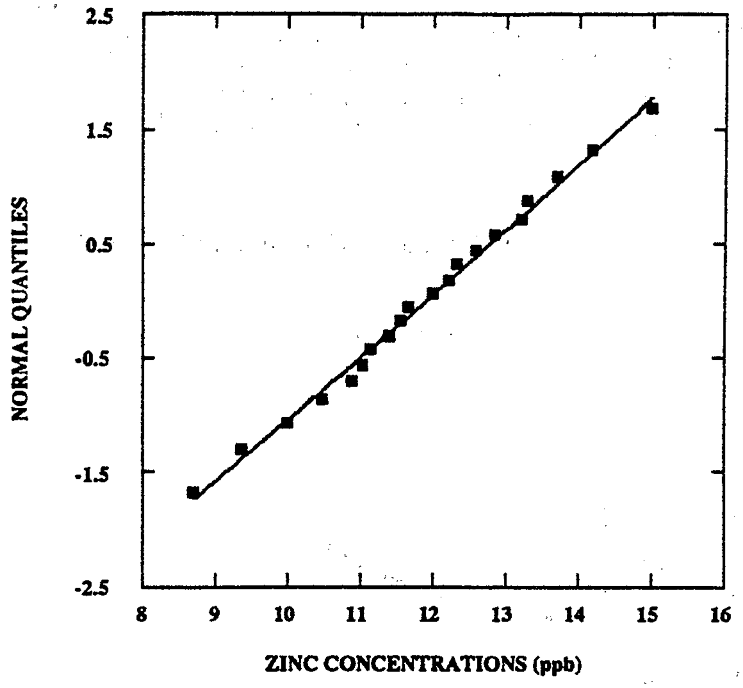
detected values: the Censored Probability Plot has a correlation of r=.969, while the Detects-Only Probability Plot has a correlation of r=.998.

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 $\overline{x}^*$ and s^* denote respectively the sample mean and standard deviation of the detected values, the adjusted overall mean can be estimated as

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

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

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

The general formula for a parametric statistical interval adjusted for nondetects by Aitchison's method is given by $\hat{\mu} \pm \kappa \cdot \hat{\sigma}$, 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 $\overline{x}^{*} = 11.891$ 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:

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

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

If Cohen's adjustment is mistakenly computed on these data instead, with a detection limit of 7 ppb,the estimates become $\hat{\mu} = 7.63$ and $\hat{\sigma} = 4.83$. 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 magnitudes 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 all of k future measurements of an analyte with confidence level 1-α, 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

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

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 sum 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 $\beta\%$ coverage (e.g., 95%) once a probable rate λ has been estimated, one must compute the Poisson percentile that is larger than $\beta\%$ of all possible measurements from that distribution, that is, the $\beta\%$ 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 $\beta\%$ 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 $(1-\beta)\%$ 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 $\beta\%$ coverage with $\gamma\%$ 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_{T_n} 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 $2\lambda_{T_n}$=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	$\overline{R}_b$=7.9		$\overline{R}_3$=12.2	$\overline{R}_4$=17.7	$\overline{R}_5$=19.3

Step 3. Calculate the sums of the ranks in each group (R_i) and the mean ranks in each group ($\overline{R}_i$). 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: $\overline{R}_3 - \overline{R}_b = 12.2\text{-}7.9 = 4.3$

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

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

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-β)%, the κ multiplier can be computed directly with the aid of a Student's t-distribution table. In this case, the formula becomes

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

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

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

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 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. 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 average (not the minimum 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 expected coverage 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

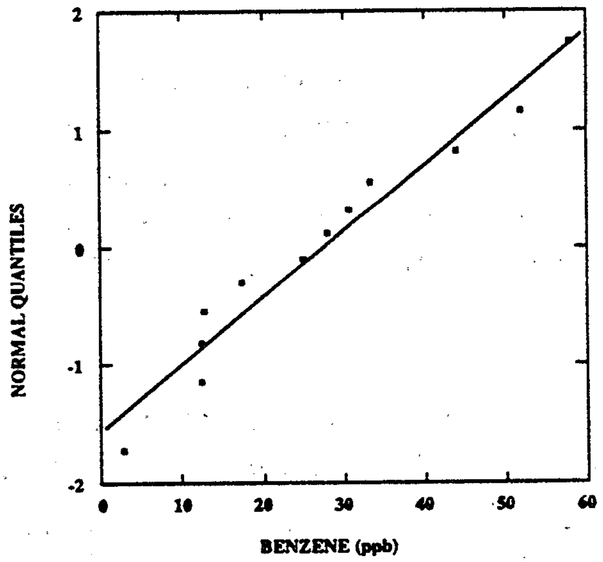
Month 3	58.1	Month 5	47.6
	12.6		3.8
	17.6		2.6
	25.3		51.9
	n=12		n=4
	Mean=27.52		Mean=26.48
	SD=17.10		SD=26.94

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 reliable 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.

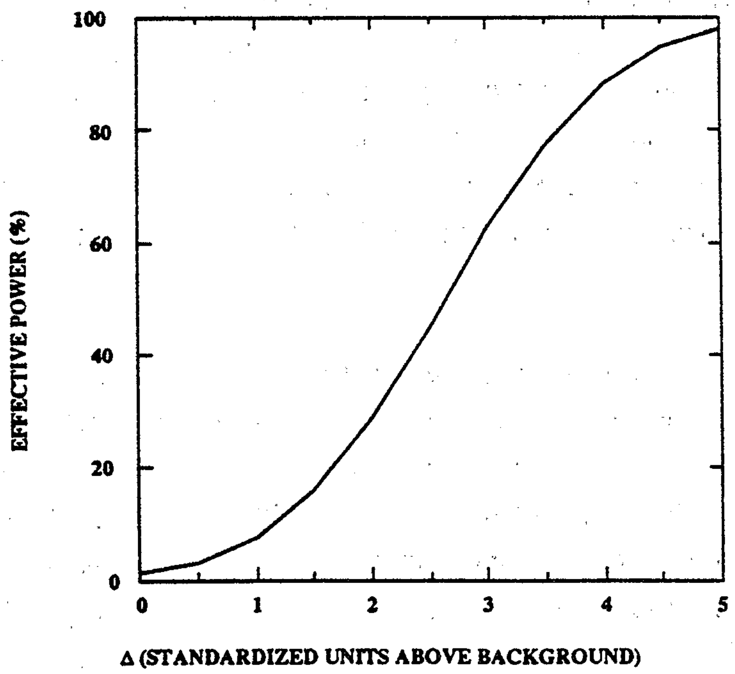
Perhaps the best way to describe the power function associated with a particular testing procedure is via a graph, such as the example below of the power of a standard Normal-based upper Prediction limit with 99% confidence. The power in percent is plotted along the y-axis against the standardized mean 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

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

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 Δ=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 Δ=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% per constituent, 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

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

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 $x=\Pr\{X_i \leq TL\}$ and $y=\Pr\{Y_i \leq PL \mid X_i > TL\}$, 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 $\Pr\{Y_i \leq PL\}$ 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
5	8	95	90	**
	16	95	90	**
	16	95	85	*
	24	95	85	**
	24	95	90	*
20	8	95	98	**
	16	95	97	**
	24	95	97	**
50	16	98	97	**
	16	99	92	*
	24	98	95	**
	24	99	90	**
100	16	98	98	*
	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
5	8	1 Resample		*
	8	2 Resamples (A)	Gibbons, 1990	**
	16	1 Resample		**
	16	2 Resamples (B)	Gibbons, 1991	**
	24	2 Resamples (B)	Gibbons, 1991	**
20	8	2 Resamples (A)	Gibbons, 1990	*
	16	1 Resample		*
	16	2 Resamples (A)	Gibbons, 1990	*
	24	1 Resample		**
	24	2 Resamples (B)	Gibbons, 1991	*
	32	1 Resample		*
	32	2 Resamples (B)	Gibbons, 1991	**
50	16	2 Resamples (A)	Gibbons, 1990	**
	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.

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

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.

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

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

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

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

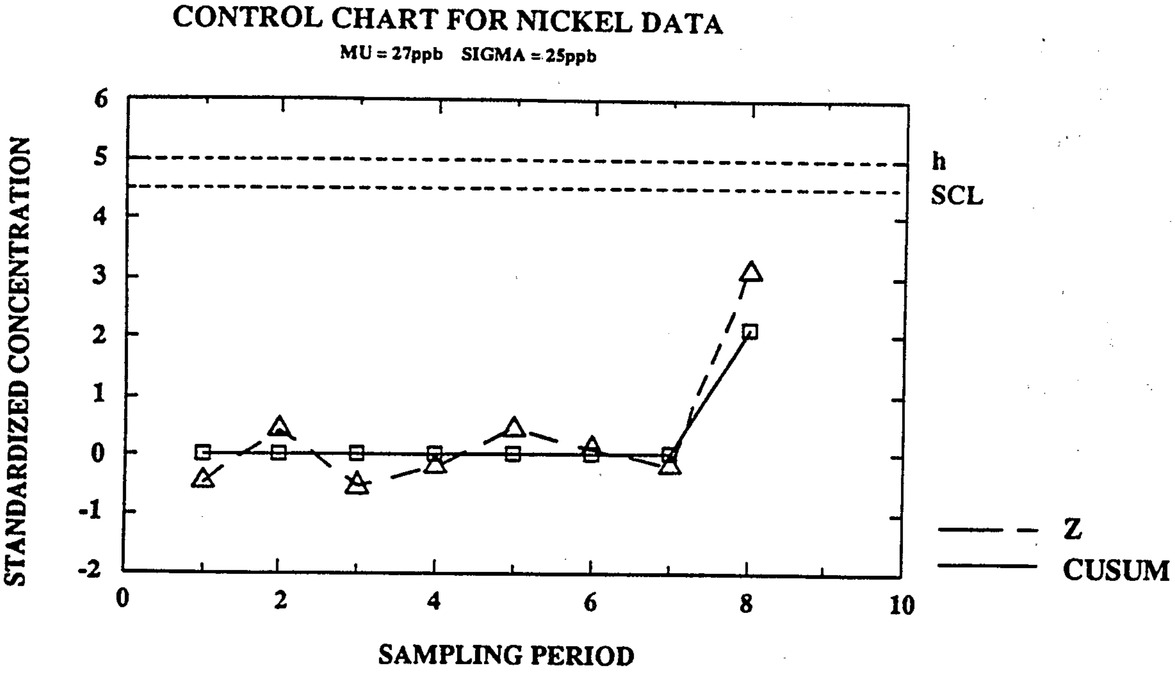
$$S_5 = \max \{0, -0.56 + 0\} = 0.00$$

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

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

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

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, <u>not</u> 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.

REFERENCES

Aitchison, J. (1955) On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of American Statistical Association*, 50(272): 901-8.

Barnett, V. and Lewis, T. (1978) Outliers in statistical data. New York: John Wiley & Sons.

Cohen, A.C., Jr. (1959) Simplified estimators for the normal distribution when samples are single censored or truncated. *Technometrics*, 1:217-37.

Cox, D.R. and Hinkley, D.V. (1974) Theoretical statistics. London: Chapman & Hall.

Davis, C.B. and McNichols, R.J. (1987) One-sided intervals for at least p of m observations from a normal population on each of r future occasions. *Technometrics*, 29(3):359-70.

Filliben, J.J. (1975) The probability plot correlation coefficient test for normality. *Technometrics*, 17:111-7.

Gan, F.F. and Koehler, K.J. (1990) Goodness-of-fit tests based on p-p probability plots. *Technometrics*, 32(3):289-303.

Gayen, A.K. (1949) The distribution of "Student's" t in random samples of any size drawn from non-normal universes. *Biometrika*, 36:353-69.

Gibbons, R.D. (1987a) Statistical prediction intervals for the evaluation of ground-water quality. *Ground Water*, 25(4):455-65.

Gibbons, R.D. (1987b) Statistical models for the analysis of volatile organic compounds in waste disposal sites. *Ground Water*, 25(5):572-80.

Gibbons, R.D. (1990) A general statistical procedure for ground-water detection monitoring at waste disposal facilities. *Ground Water*, 28(2):235-43.

Gibbons, R.D. (1991a) Statistical tolerance limits for ground-water monitoring. *Ground Water*, 29(4):563-70.

Gibbons, R.D. (1991b) Some additional nonparametric prediction limits for ground-water detection monitoring at waste disposal facilities. *Ground Water*, 29(5):729-36.

Gilliom, R.J. and Helsel, D.R. (1986) Estimation of distributional parameters for censored trace level water quality data: part 1, estimation techniques. *Water Resources Research*, 22(2):135-46.

Guttman, I. (1970) Statistical tolerance regions: classical and bayesian. Darien, Connecticut: Hafner Publishing.

Hahn, G.J. (1970) Statistical intervals for a normal population: part 1, tables, examples, and applications. *Journal of Quality Technology*, 2(3):115-25.

Lehmann, E.L. (1975) Nonparametrics: statistical methods based on ranks. San Francisco: Holden-Day, Inc.

Madansky, A. (1988) Prescriptions for working statisticians. New York: Springer-Verlag.

McBean, E.A. and Rovers, F.A. (1992) Estimation of the probability of exceedance of contaminant concentrations. *Ground Water Monitoring Review*, Winter, 115-9.

McNichols, R.J. and Davis, C.B. (1988) Statistical issues and problems in ground water detection monitoring at hazardous waste facilities. *Ground Water Monitoring Review*, Fall.

Miller, R.G., Jr. (1986) Beyond ANOVA, basics of applied statistics. New York: John Wiley & Sons.

Milliken, G.A. and Johnson, D.E. (1984) Analysis of messy data: volume 1, designed experiments. Belmont, California: Lifetime Learning Publications.

Ott, W.R. (1990) A physical explanation of the lognormality of pollutant concentrations. *Journal of Air Waste Management Association*, 40:1378-83.

Ryan, T.A., Jr. and Joiner, B.L. (1990) Normal probability plots and tests for normality. *Minitab Statistical Software: Technical Reports*, November, 1–1 to 1–14.

Shapiro, S.S. and Wilk, M.B. (1965) An analysis of variance test for normality (complete samples). *Biometrika*, 52:591-611.

Shapiro, S.S. and Francia, R.S. (1972) An approximate analysis of variance test for normality. *Journal of American Statistical Association*, 67(337):215-6.

Zacks, S. (1970) Uniformly most accurate upper tolerance limits for monotone likelihood ratio families of discrete distributions. *Journal of American Statistical Association*, 65(329):307-16.

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 $N \geq 35$

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

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).

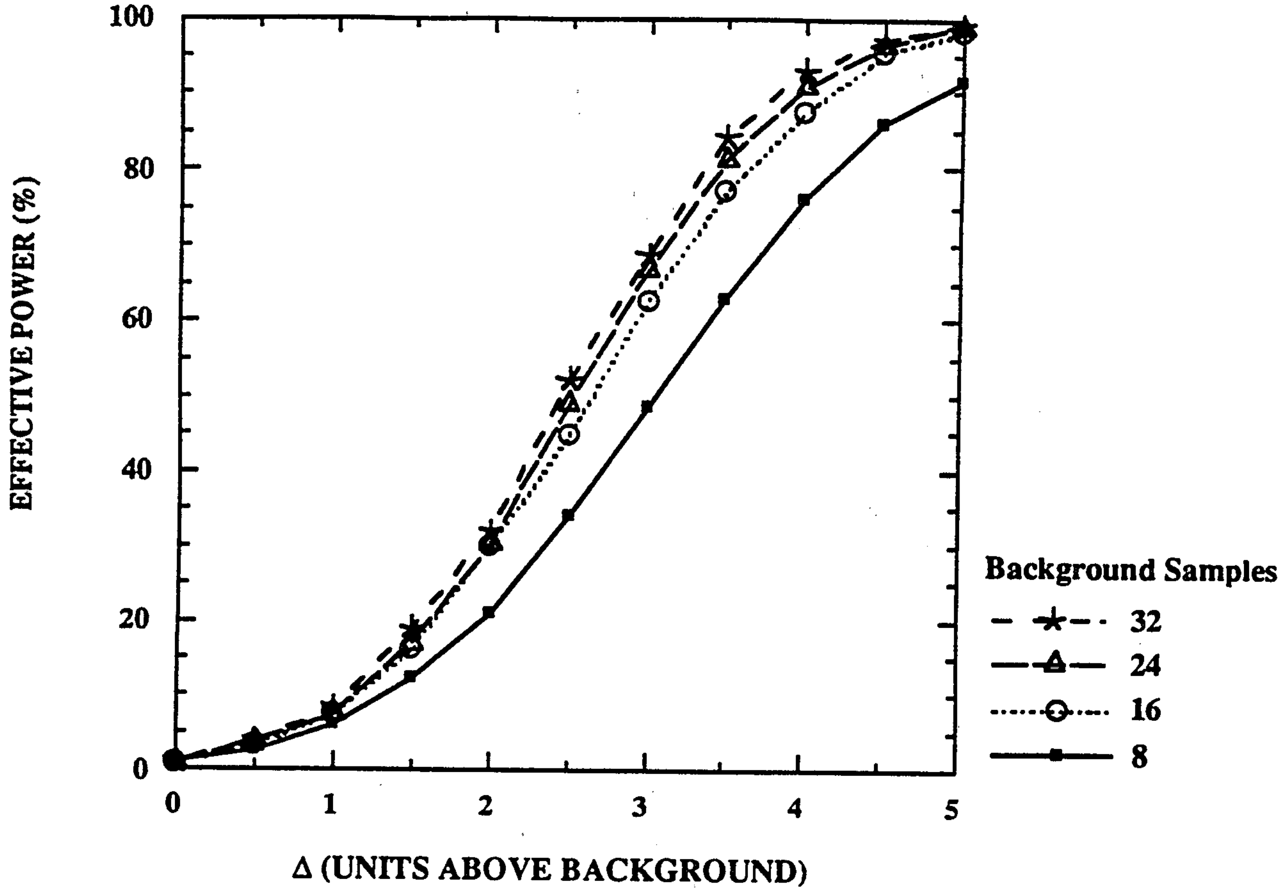
Power values for the EPA Reference Power Curves were not simulated, but represent analytical calculations based on the non-central t-distribution with non-centrality parameter Δ. SAS programs for simulating the effective power of any of the parametric or non-parametric retesting strategies are presented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

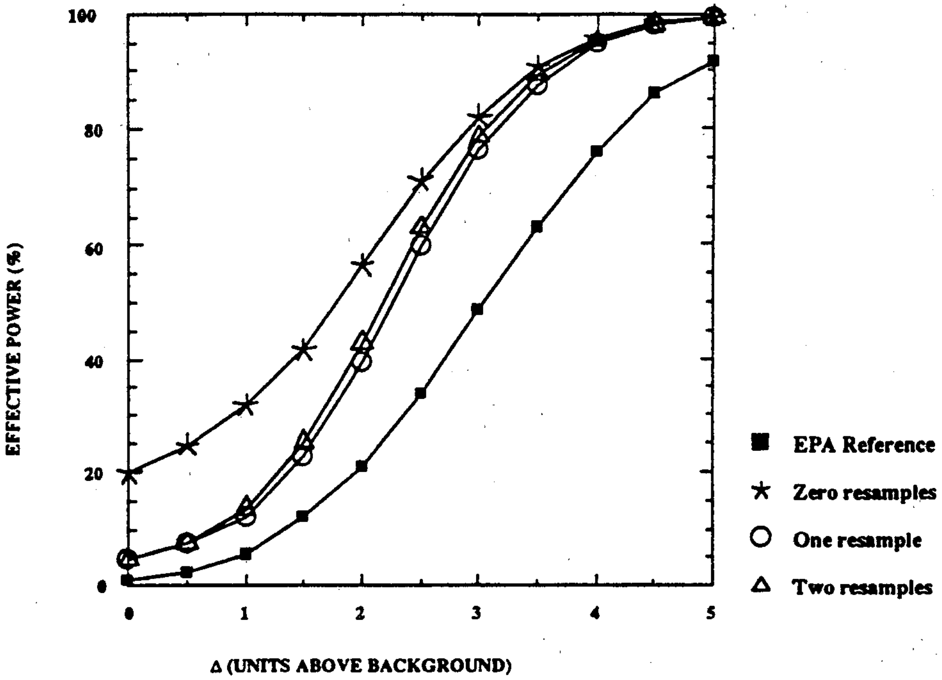
```
%LET BG1 =8;      %LET WL1 =5;
%LET BG2 =16;     %LET WL2 =5;
%LET BG3 =24;     %LET WL3 =5;
%LET BG4 =8;      %LET WL4 =20;
%LET BG5 =16;     %LET WL5 =20;
%LET BG6 =24;     %LET WL6 =20;
%LET BG7 =8;      %LET WL7 =50;
%LET BG8 =16;     %LET WL8 =50;
%LET BG9 =24;     %LET WL9 =50;
%LET BG10=8;      %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



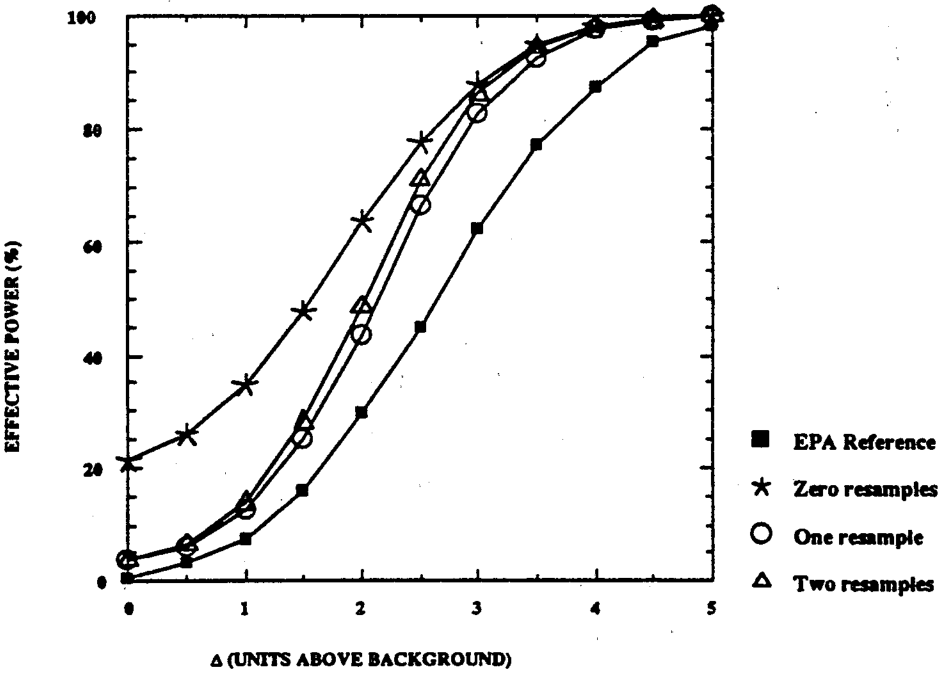
POWER CURVE FOR 95% TOLERANCE AND 90% PREDICTION LIMIT

(8 Background Samples; 5 wells)



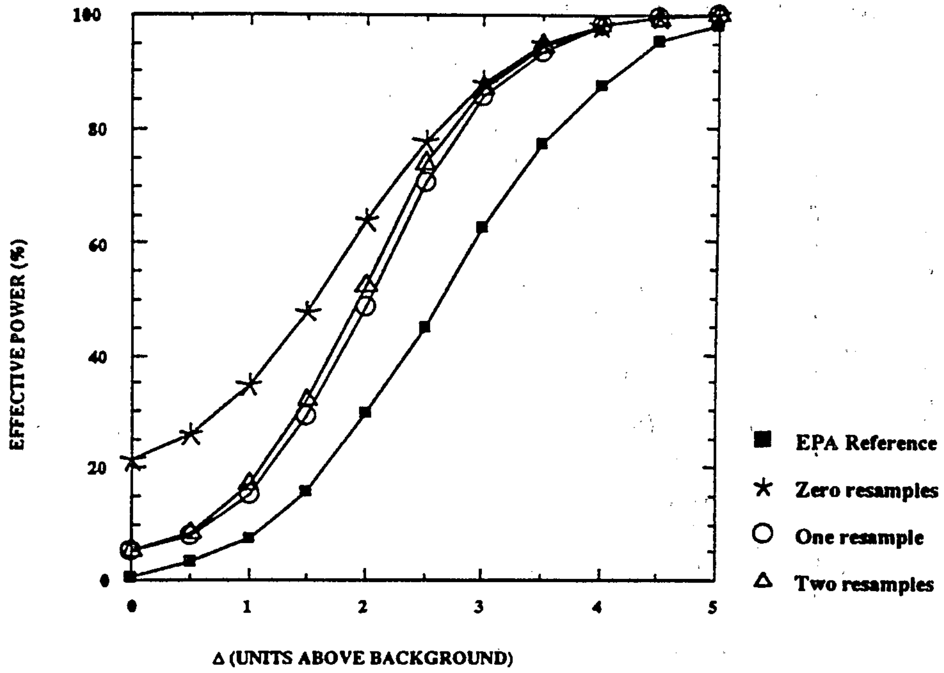
POWER CURVE FOR 95% TOLERANCE AND 90% PREDICTION LIMIT

(16 Background Samples; 5 wells)



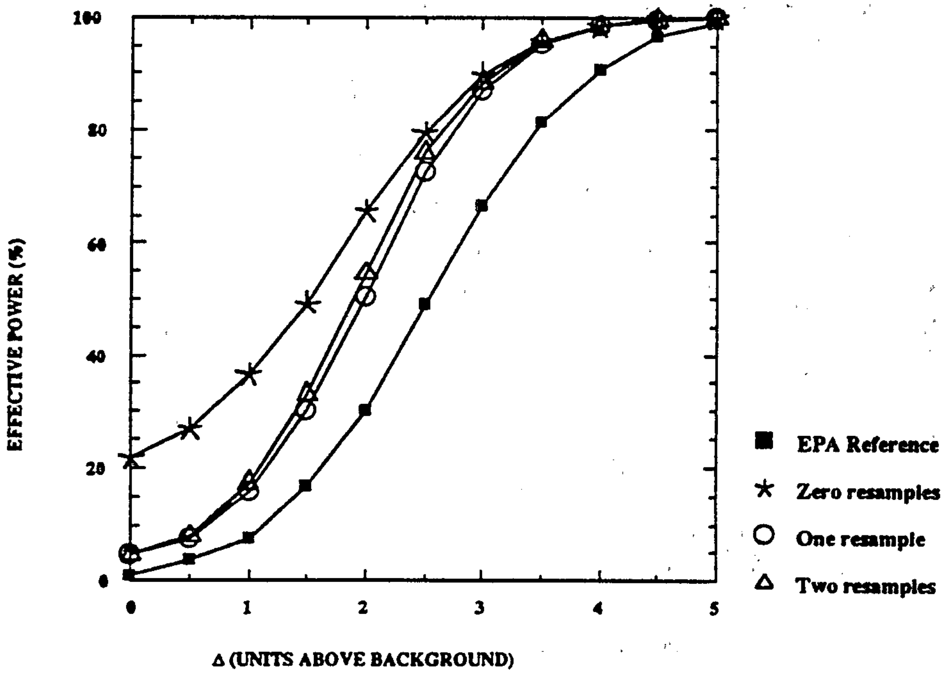
POWER CURVE FOR 95% TOLERANCE AND 85% PREDICTION LIMIT

(16 Background Samples; 5 wells)



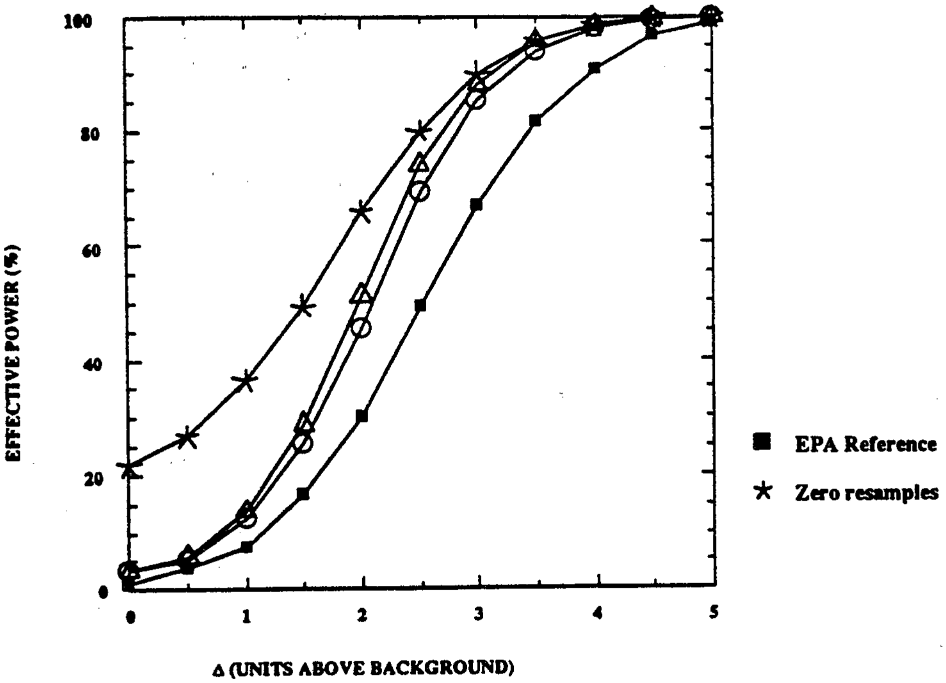
POWER CURVE FOR 95% TOLERANCE AND 85% PREDICTION LIMIT

(24 Background Samples; 5 wells)



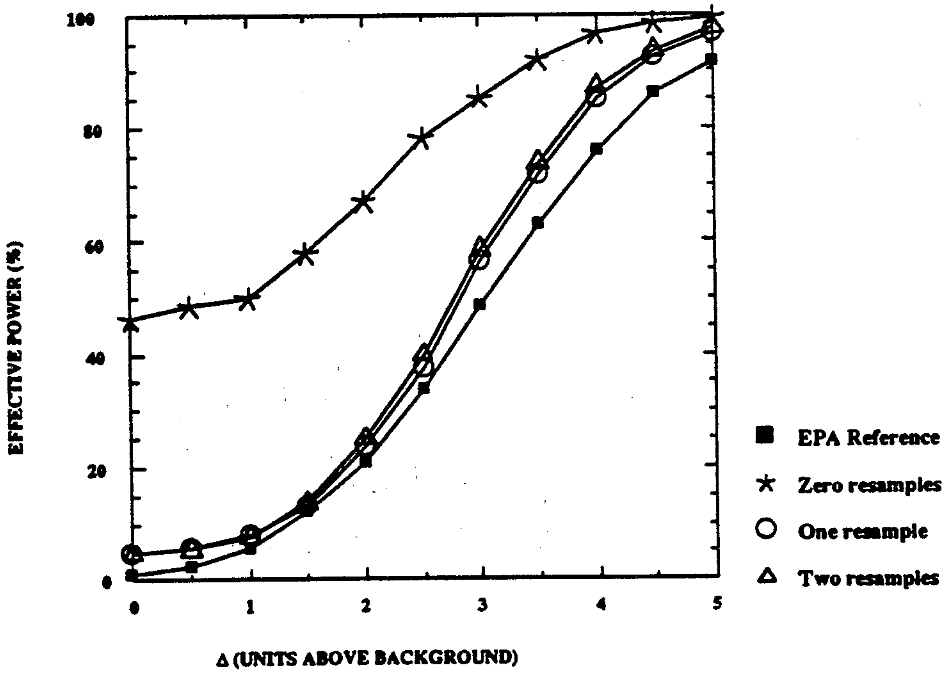
POWER CURVE FOR 95% TOLERANCE AND 90% PREDICTION LIMIT

(24 Background Samples; 5 wells)



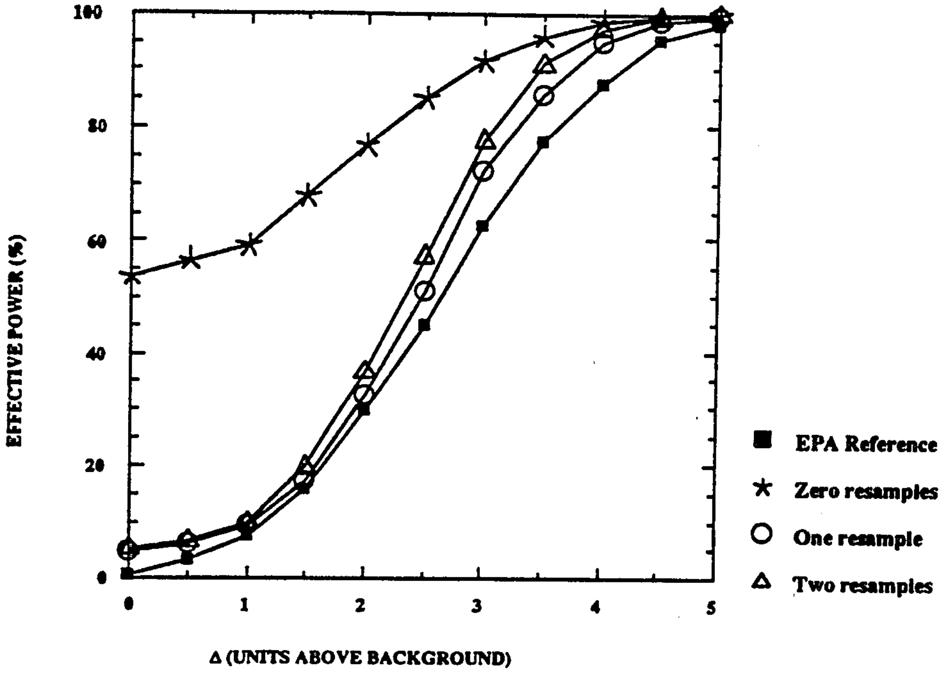
POWER CURVE FOR 95% TOLERANCE AND 98% PREDICTION LIMIT

(8 Background Samples; 20 wells)



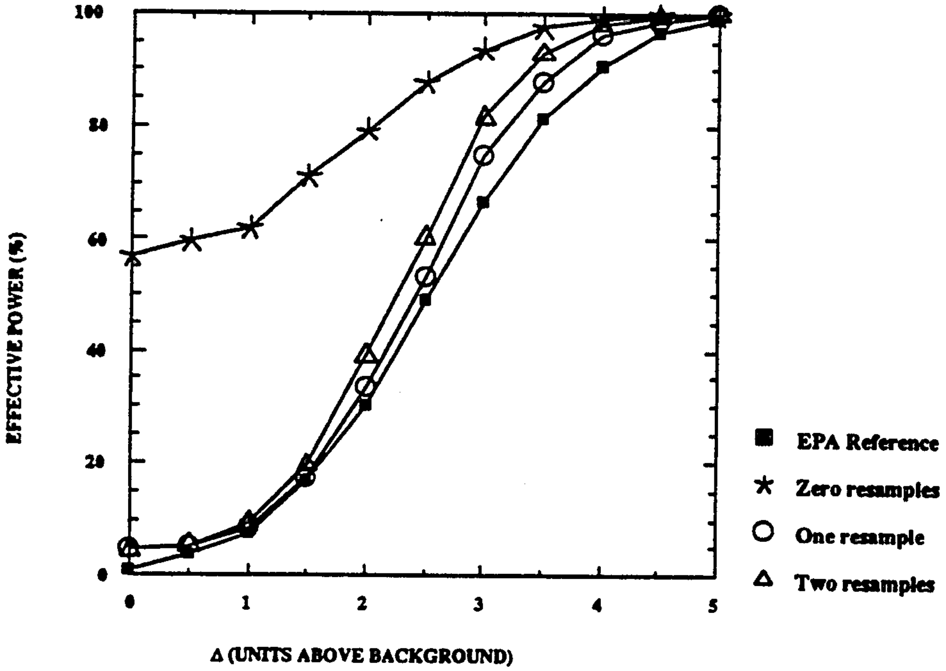
POWER CURVE FOR 95% TOLERANCE AND 97% PREDICTION LIMIT

(16 Background Samples; 20 wells)



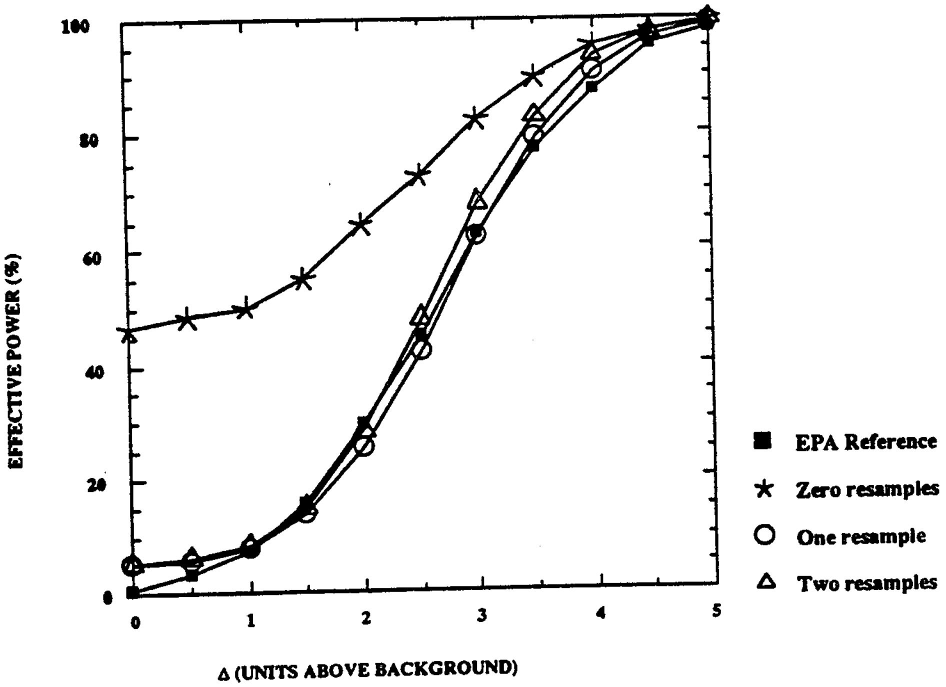
POWER CURVE FOR 95% TOLERANCE AND 97% PREDICTION LIMIT

(24 Background Samples; 20 wells)



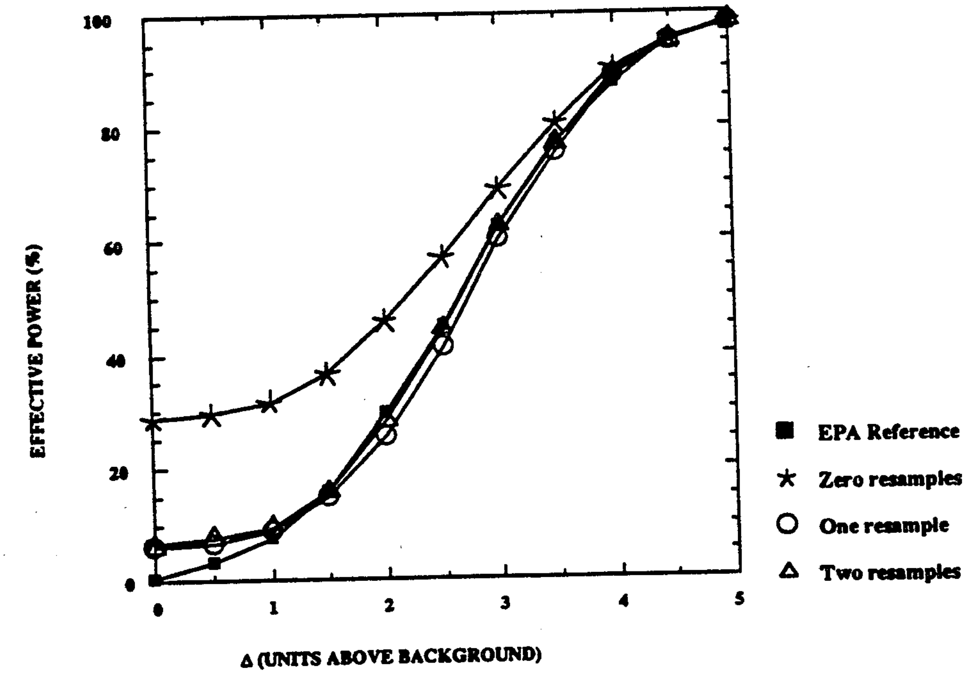
POWER CURVE FOR 98% TOLERANCE AND 97% PREDICTION LIMIT

(16 Background Samples; 50 wells)



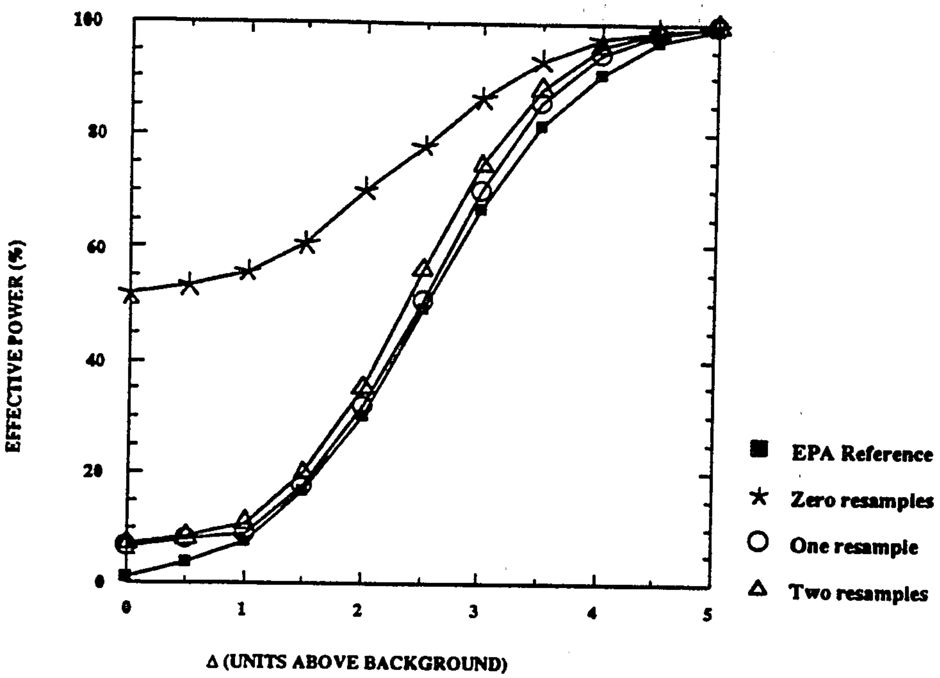
POWER CURVE FOR 99% TOLERANCE AND 92% PREDICTION LIMIT

(16 Background Samples; 50 wells)



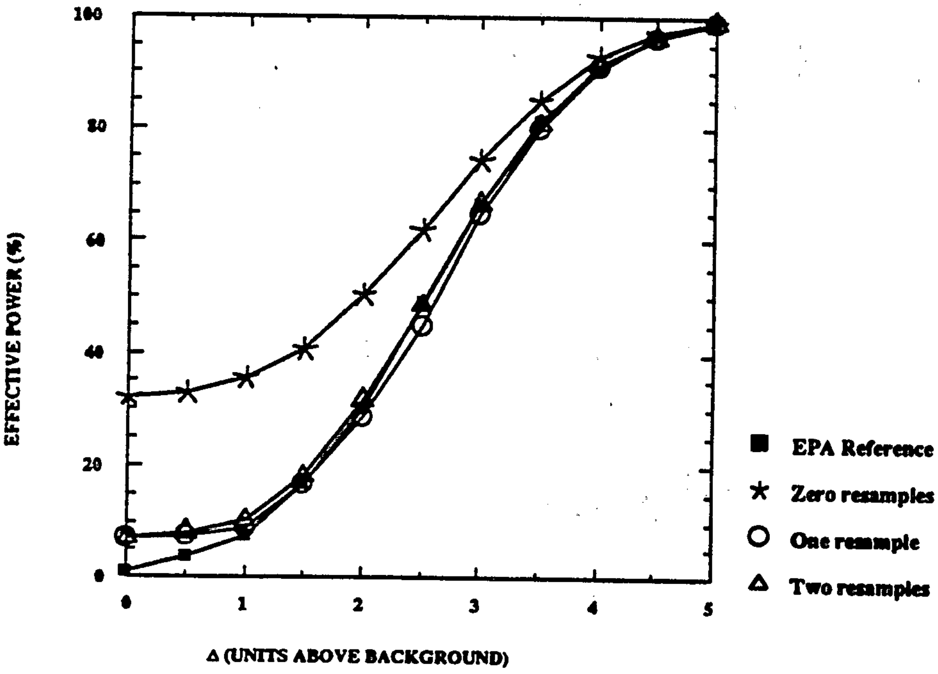
POWER CURVE FOR 98% TOLERANCE AND 95% PREDICTION LIMIT

(24 Background Samples; 50 wells)



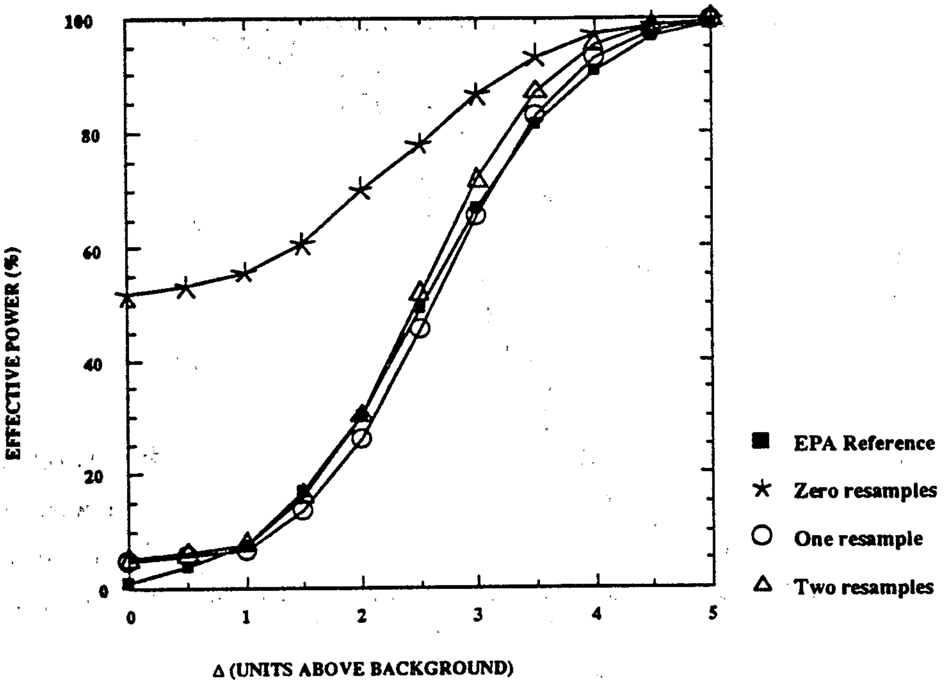
POWER CURVE FOR 99% TOLERANCE AND 90% PREDICTION LIMIT

(24 Background Samples; 50 wells)



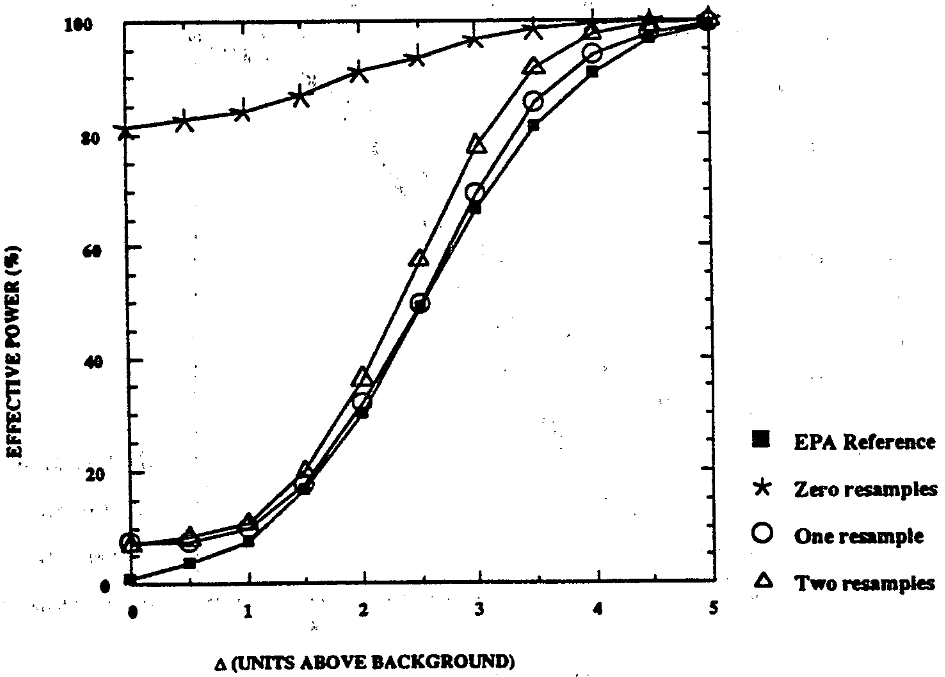
POWER CURVE FOR 98% TOLERANCE AND 97% PREDICTION LIMIT

(24 Background Samples; 50 wells)



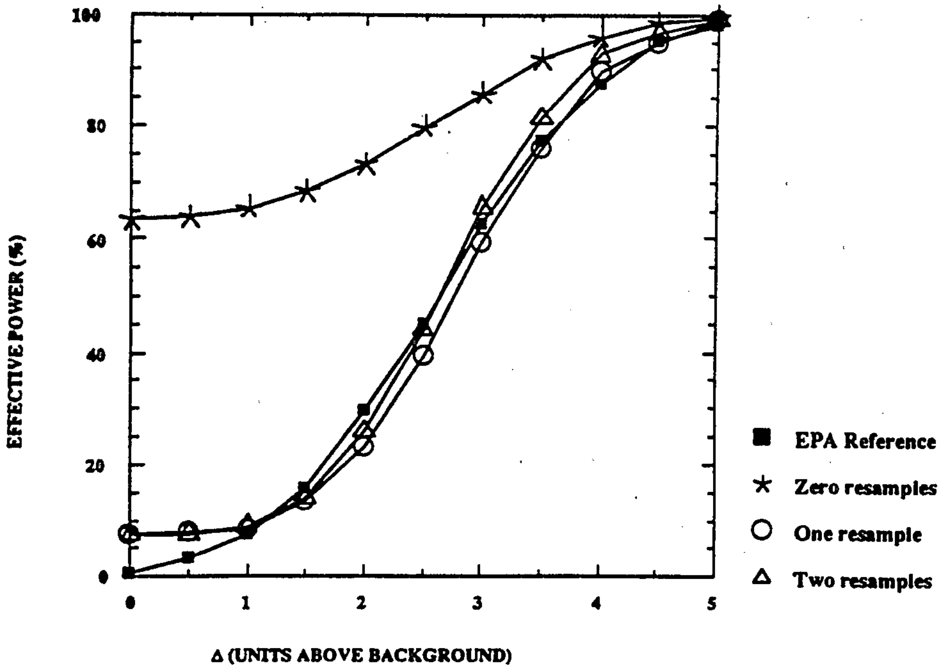
POWER CURVE FOR 95% TOLERANCE AND 98% PREDICTION LIMIT

(24 Background Samples; 50 wells)



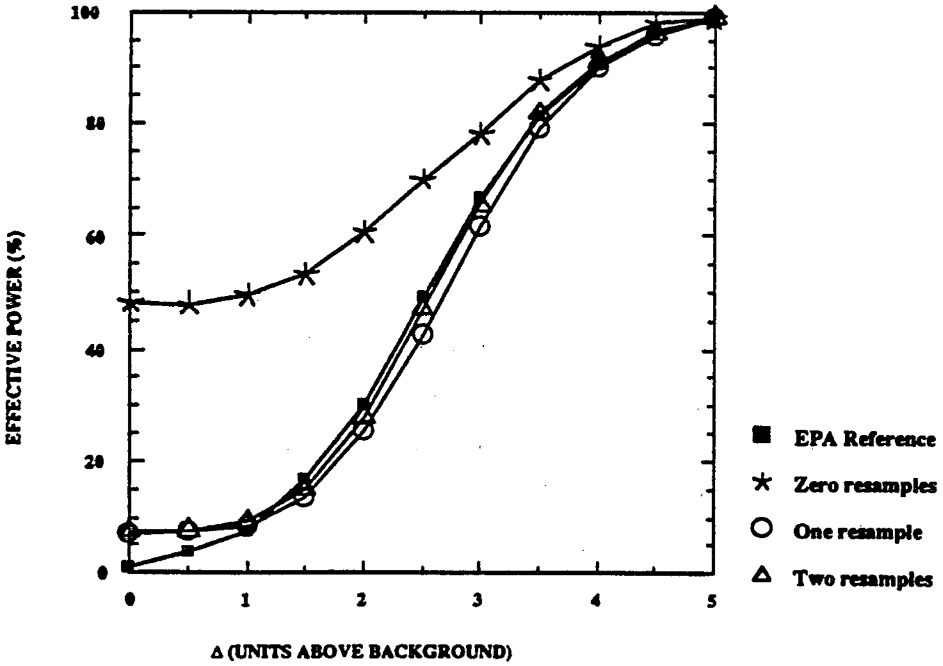
POWER CURVE FOR 98% TOLERANCE AND 98% PREDICTION LIMIT

(16 Background Samples; 100 wells)



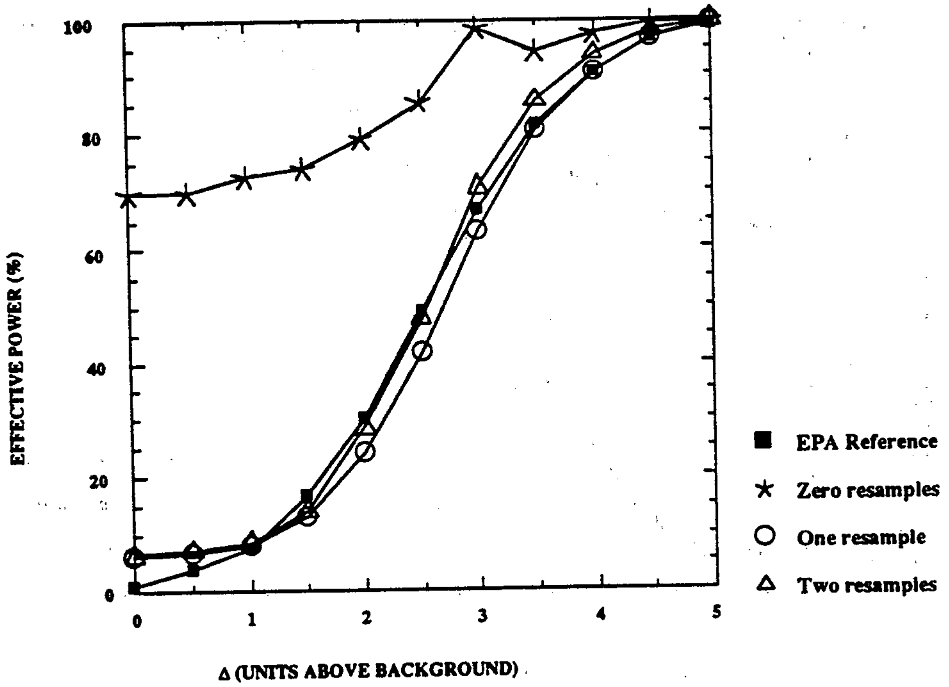
POWER CURVE FOR 99% TOLERANCE AND 95% PREDICTION LIMIT

(24 Background Samples; 100 wells)



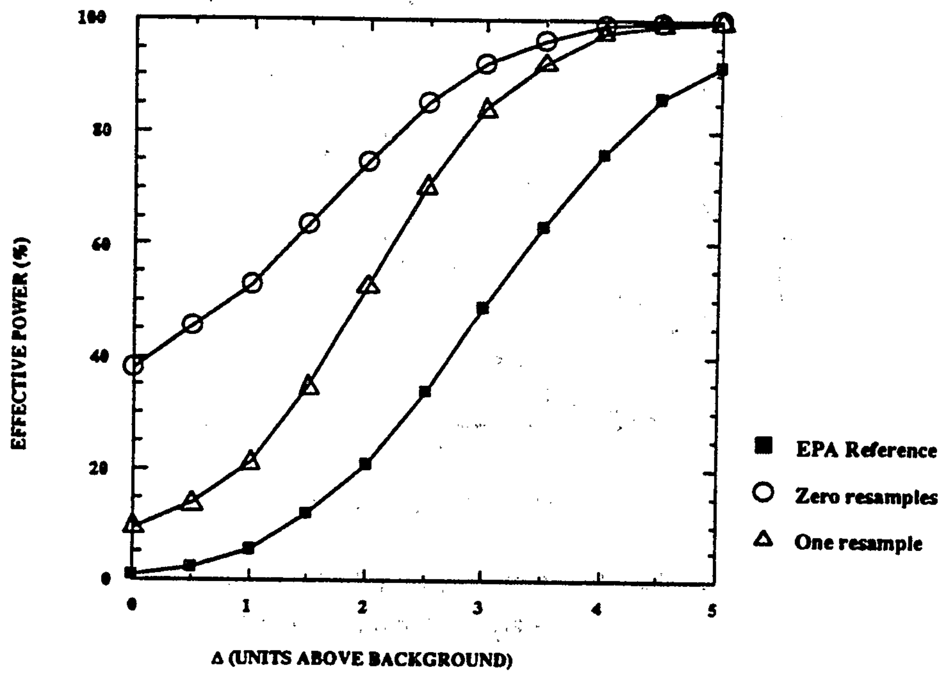
POWER CURVE FOR 98% TOLERANCE AND 98% PREDICTION LIMIT

(24 Background Samples; 100 wells)

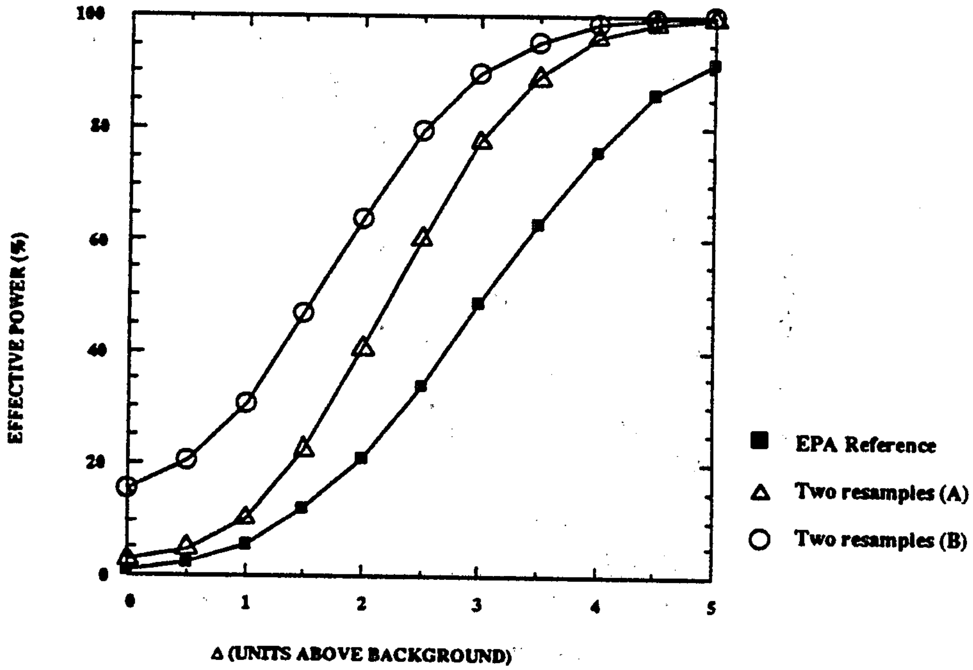


POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

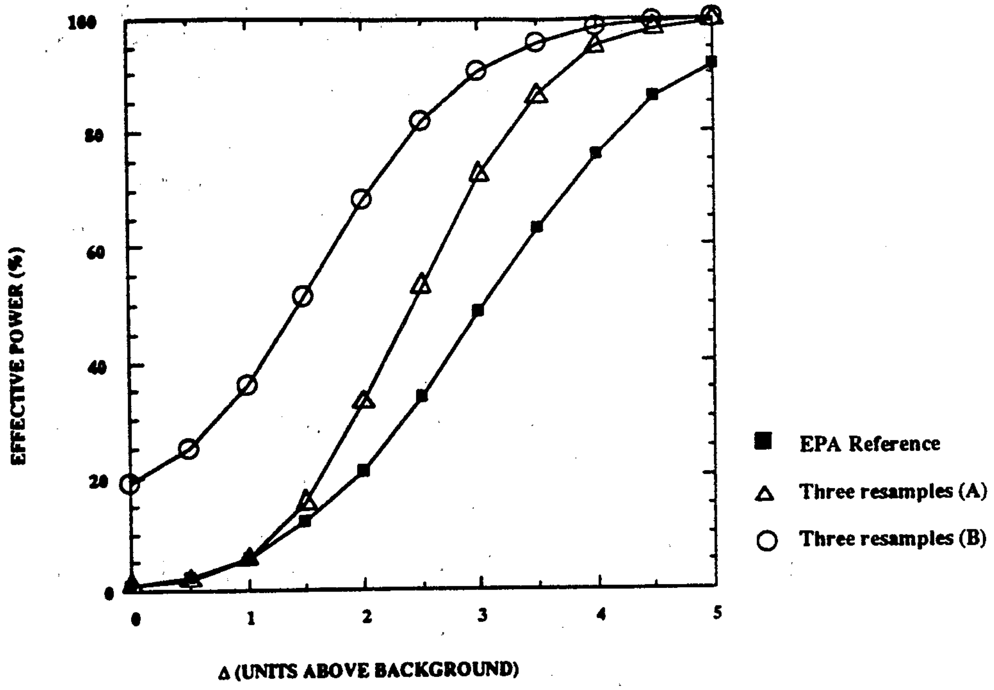
(8 Background Samples; 5 wells)



(8 Background Samples; 5 wells)

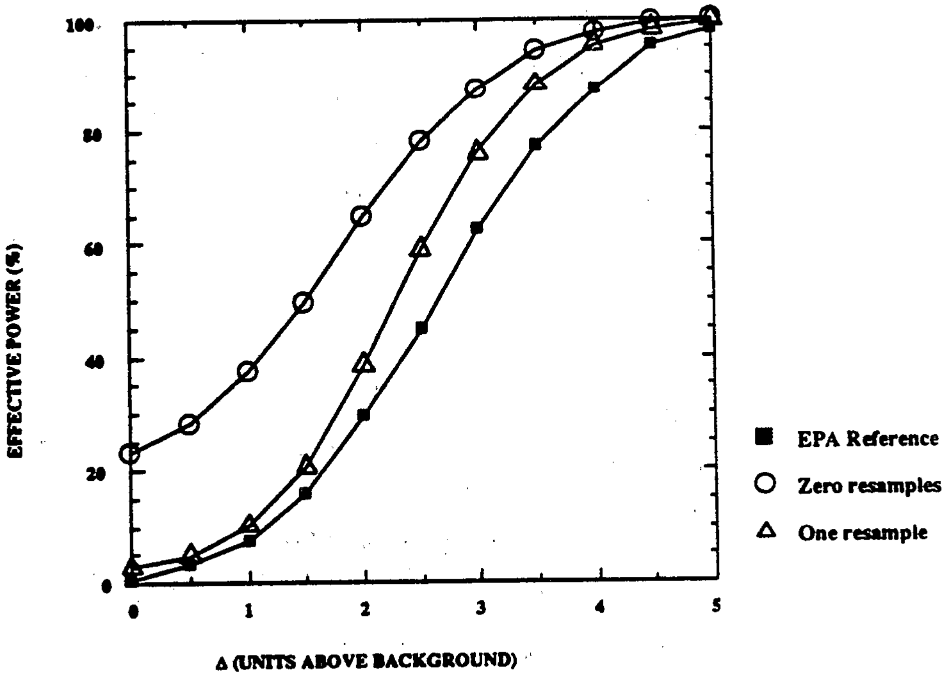


(8 Background Samples; 5 wells)

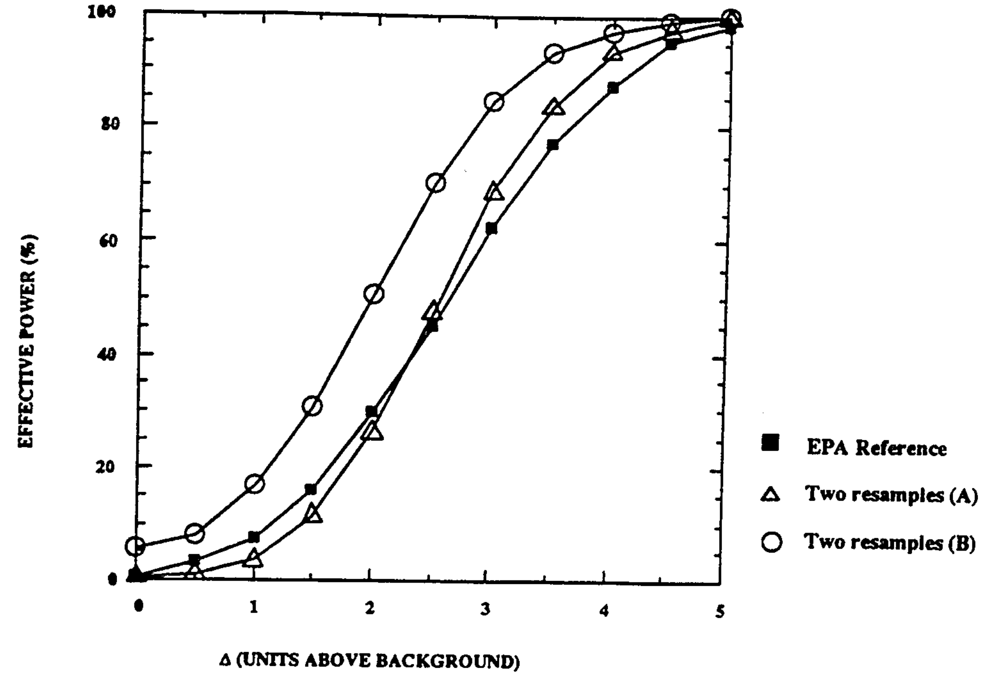


POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

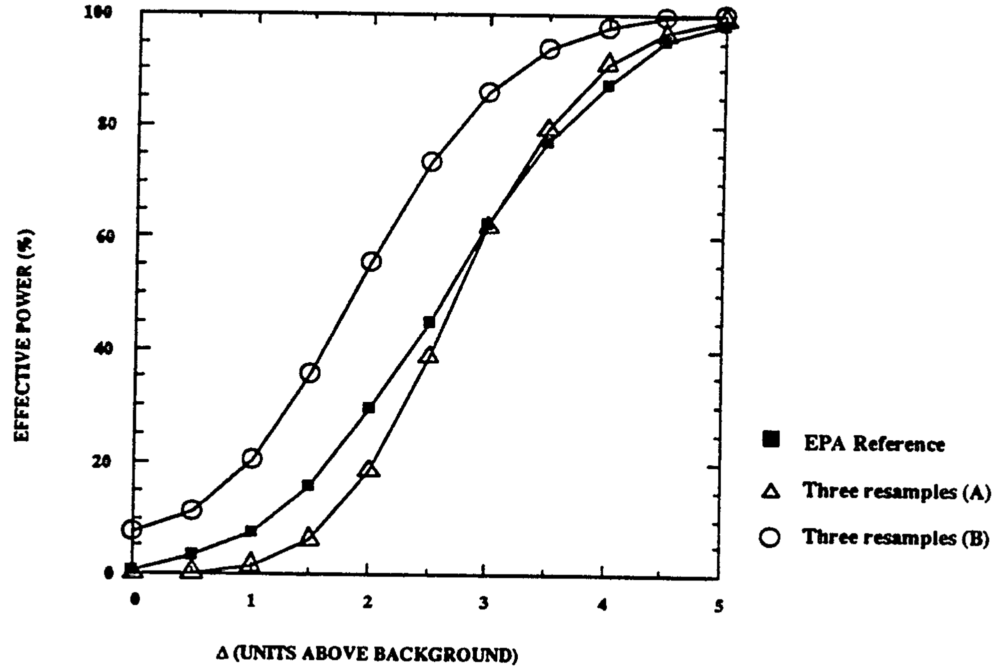
(16 Background Samples; 5 wells)



(16 Background Samples; 5 wells)

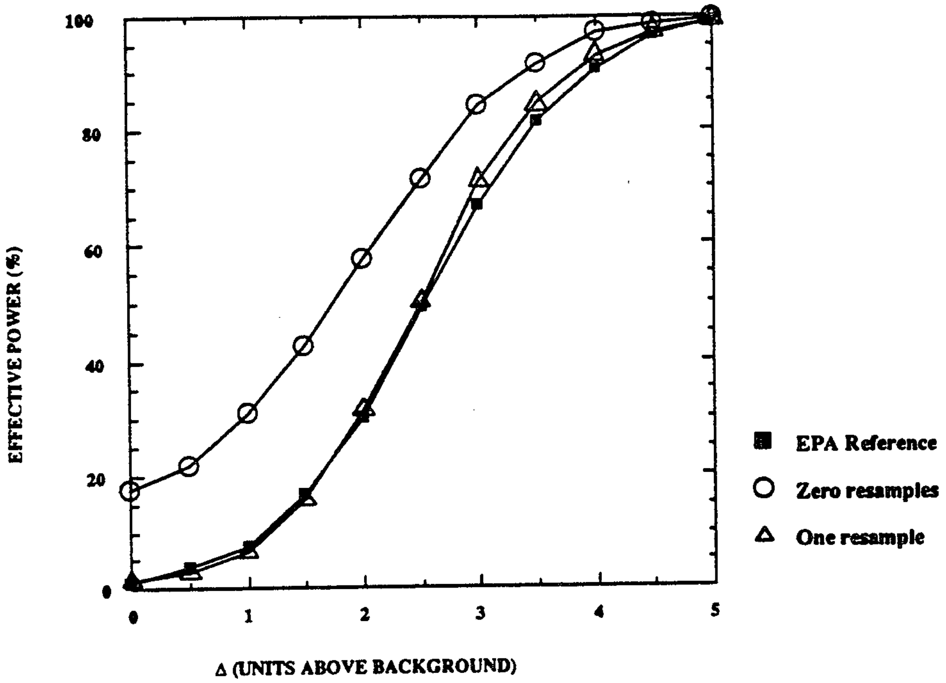


(16 Background Samples; 5 wells)

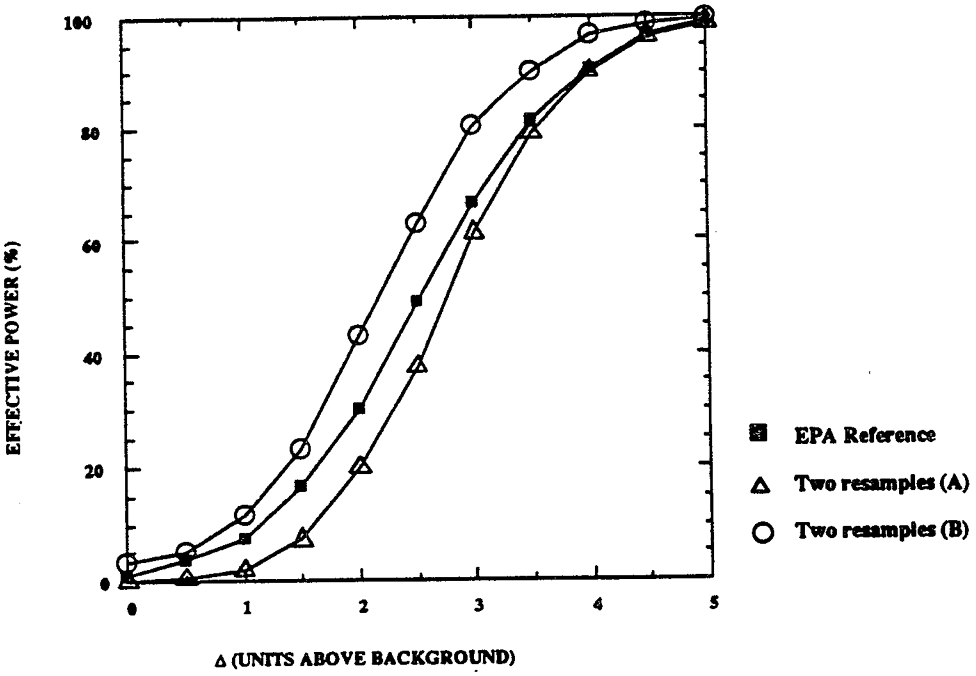


POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

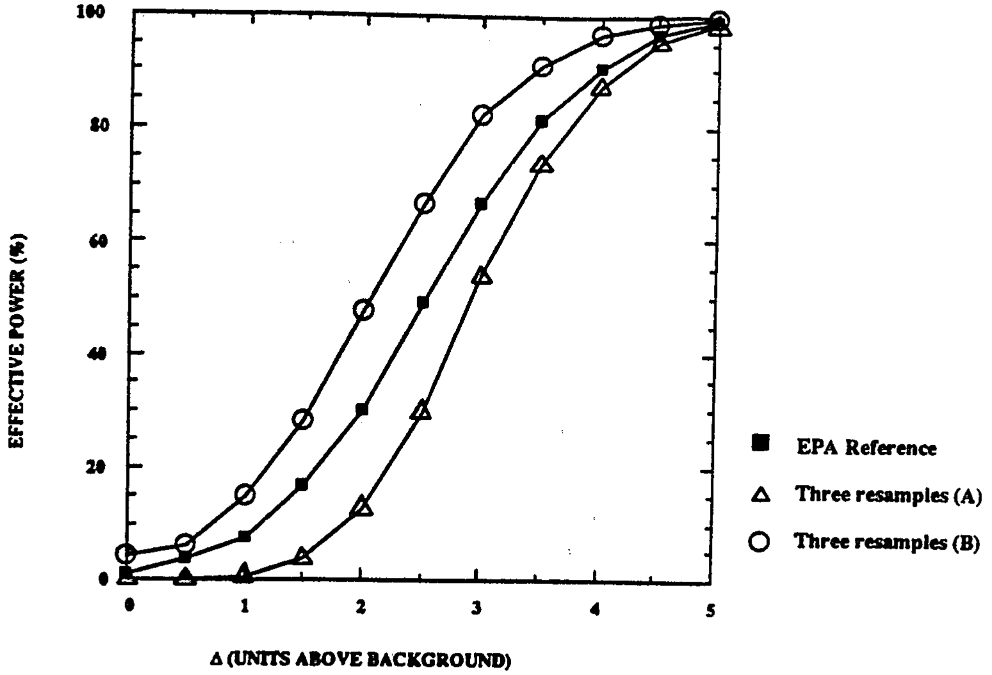
(24 Background Samples; 5 wells)



(24 Background Samples; 5 wells)

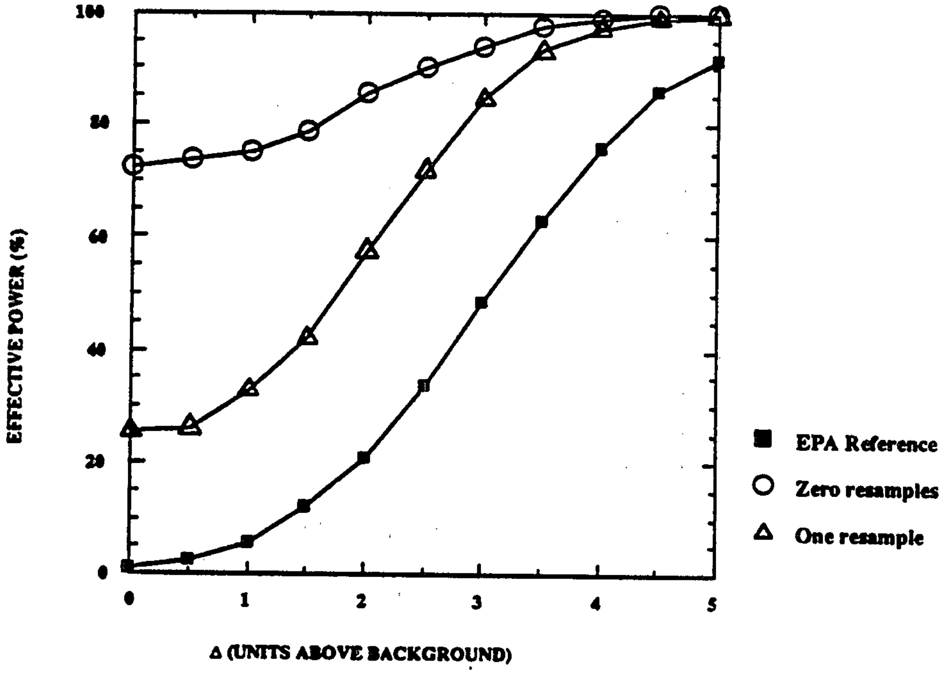


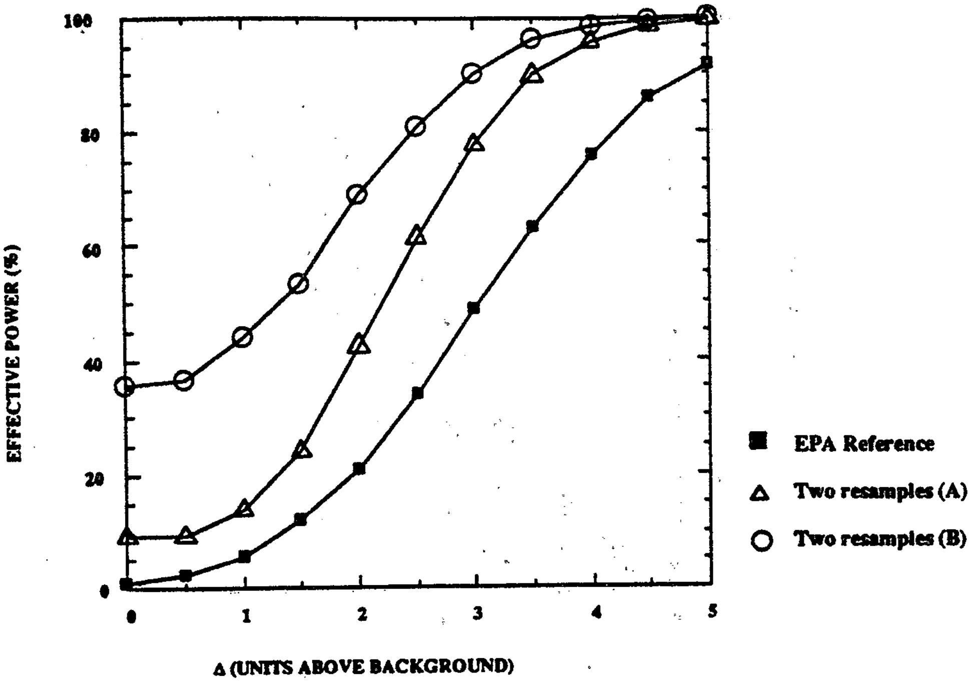
(24 Background Samples; 5 wells)



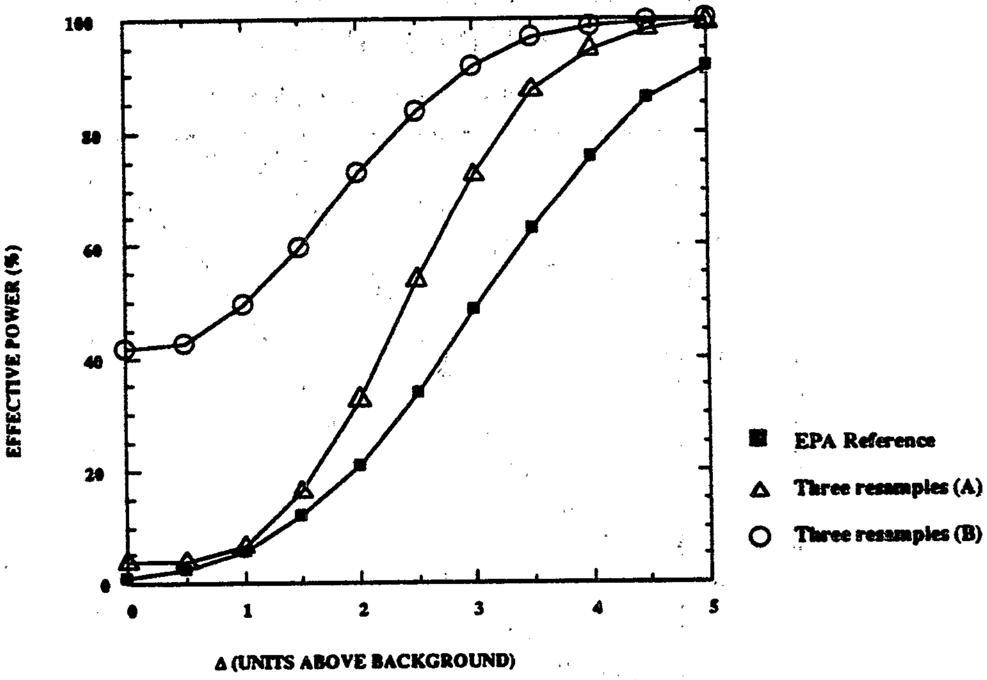
POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

(8 Background Samples; 20 wells)





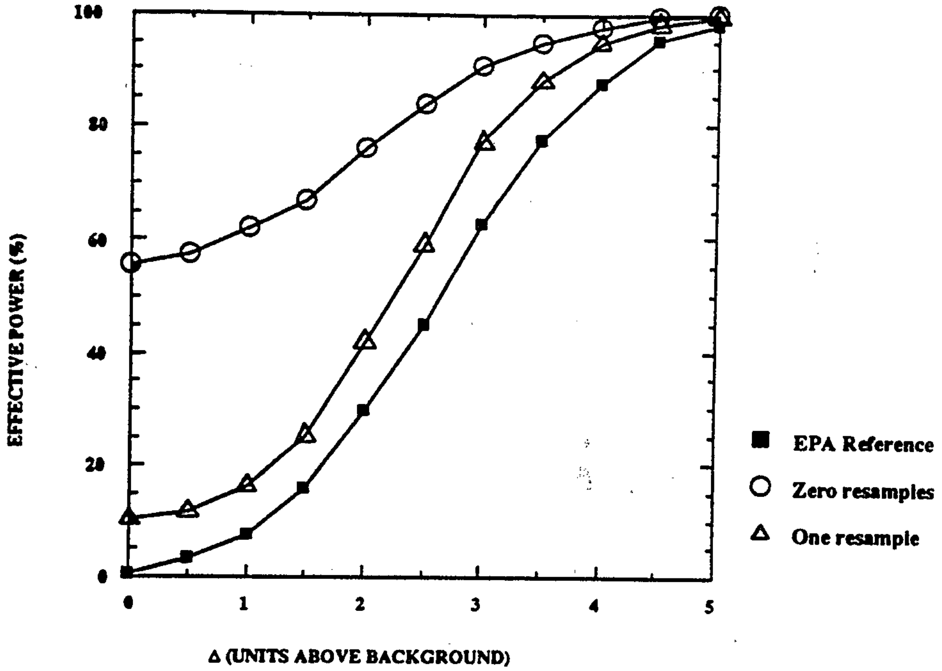
(8 Background Samples; 20 wells)



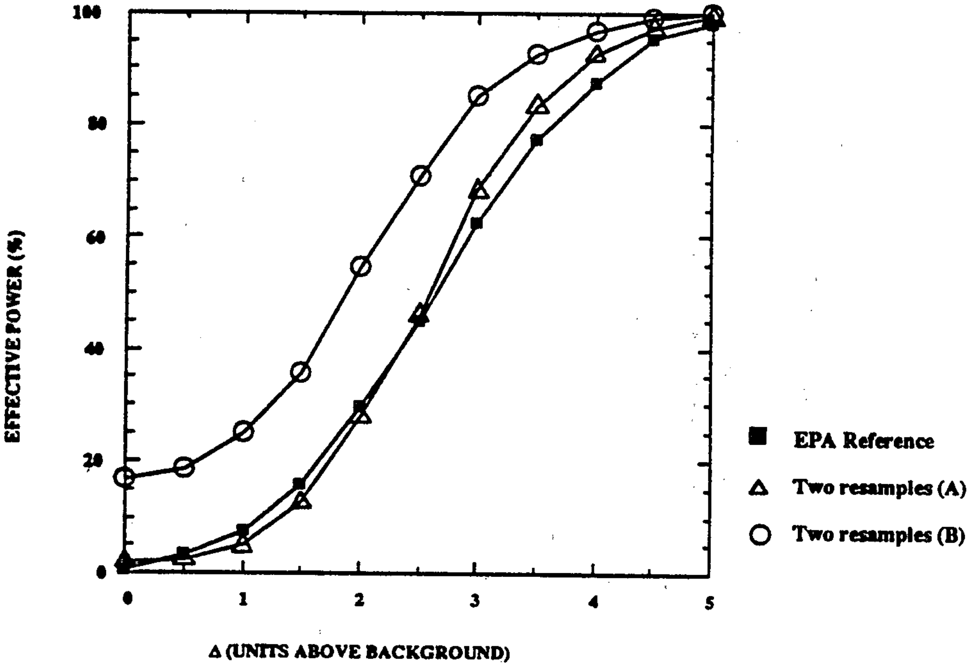
(8 Background Samples; 20 wells)

POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

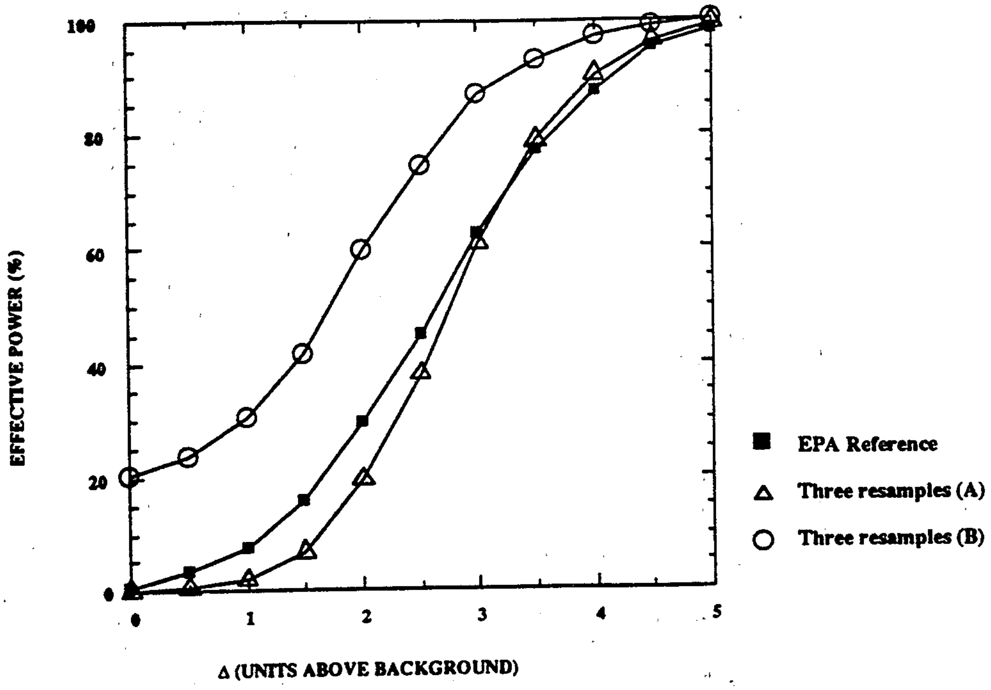
(16 Background Samples; 20 wells)



(16 Background Samples; 20 wells)

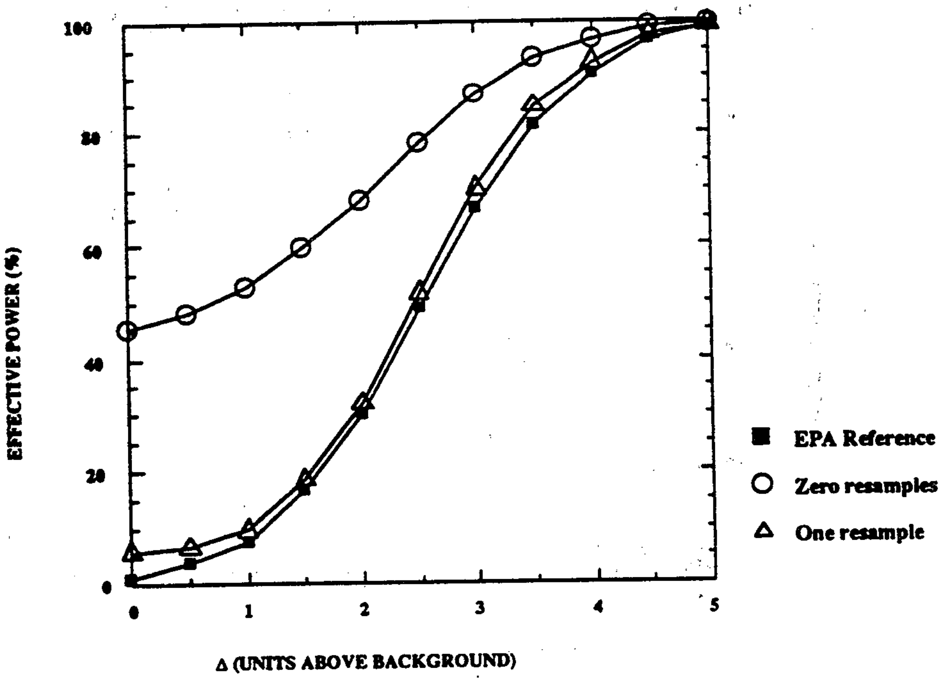


(16 Background Samples; 20 wells)

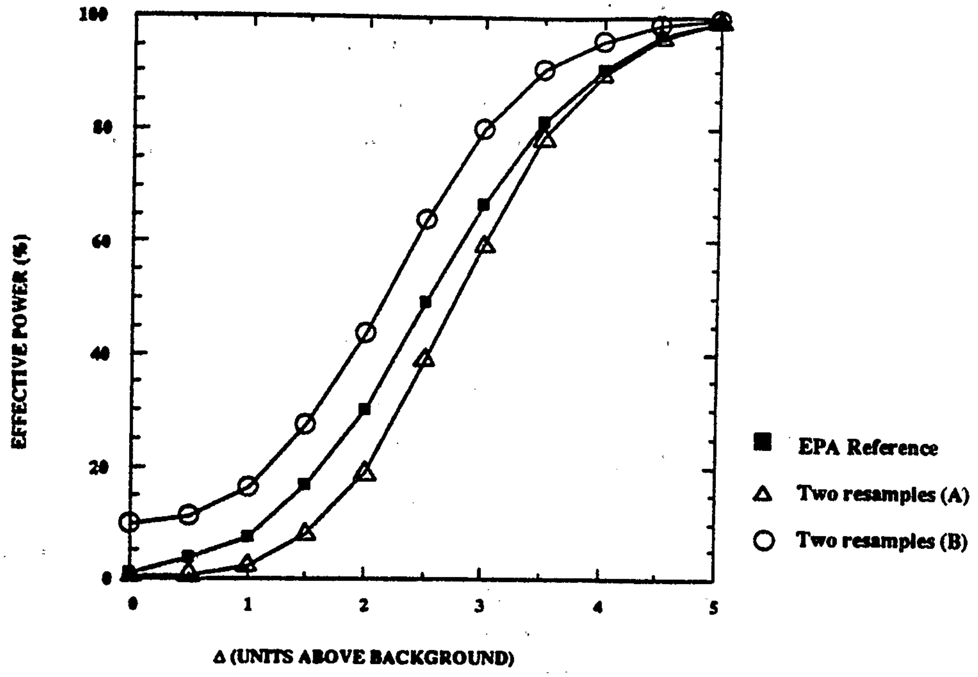


POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

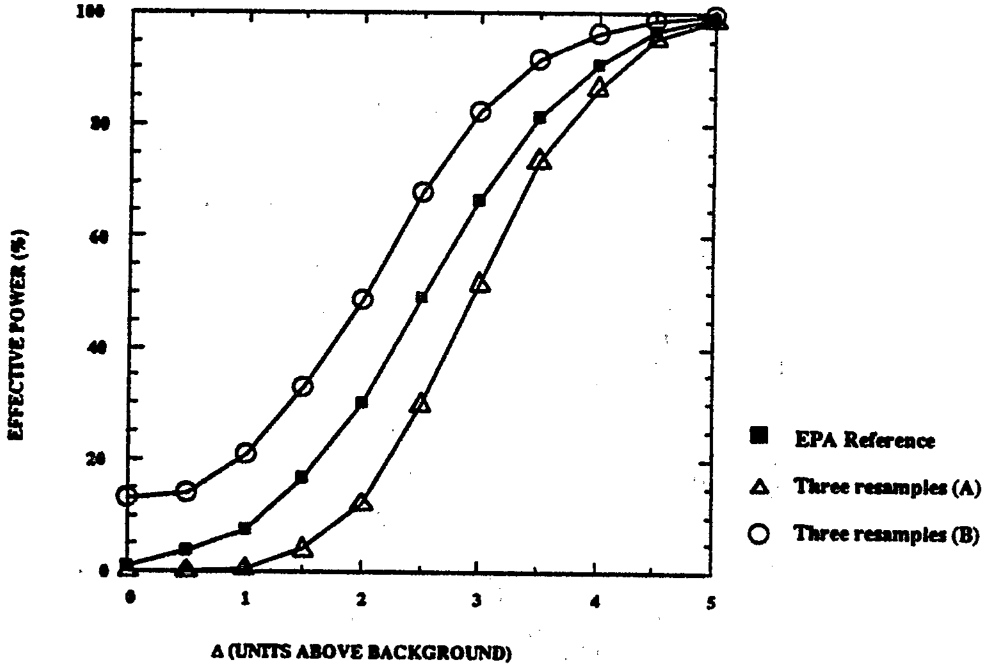
(24 Background Samples; 20 wells)



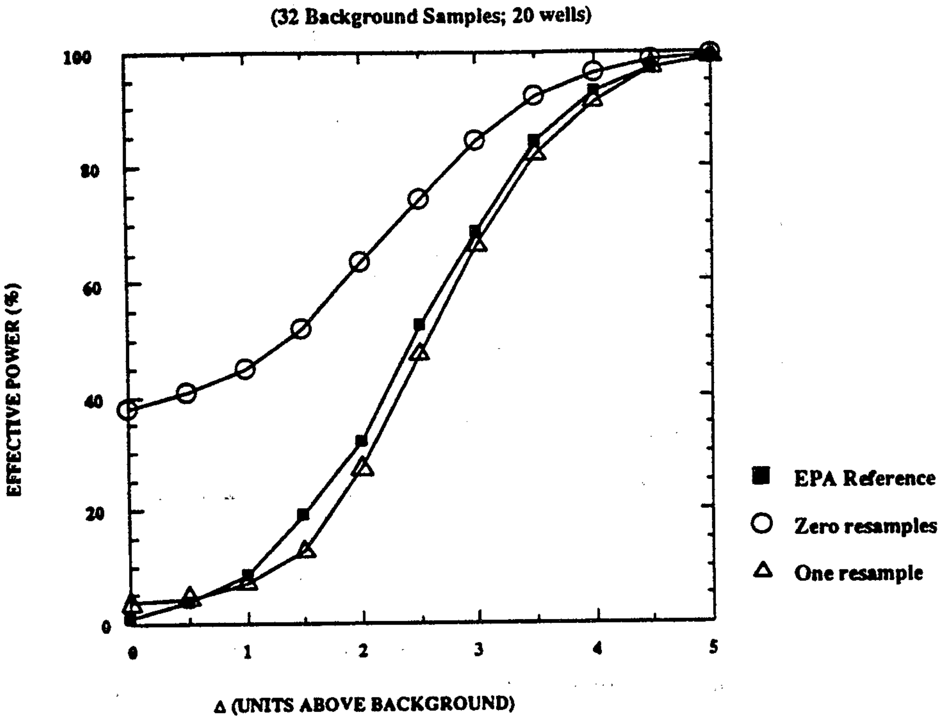
(24 Background Samples; 20 wells)

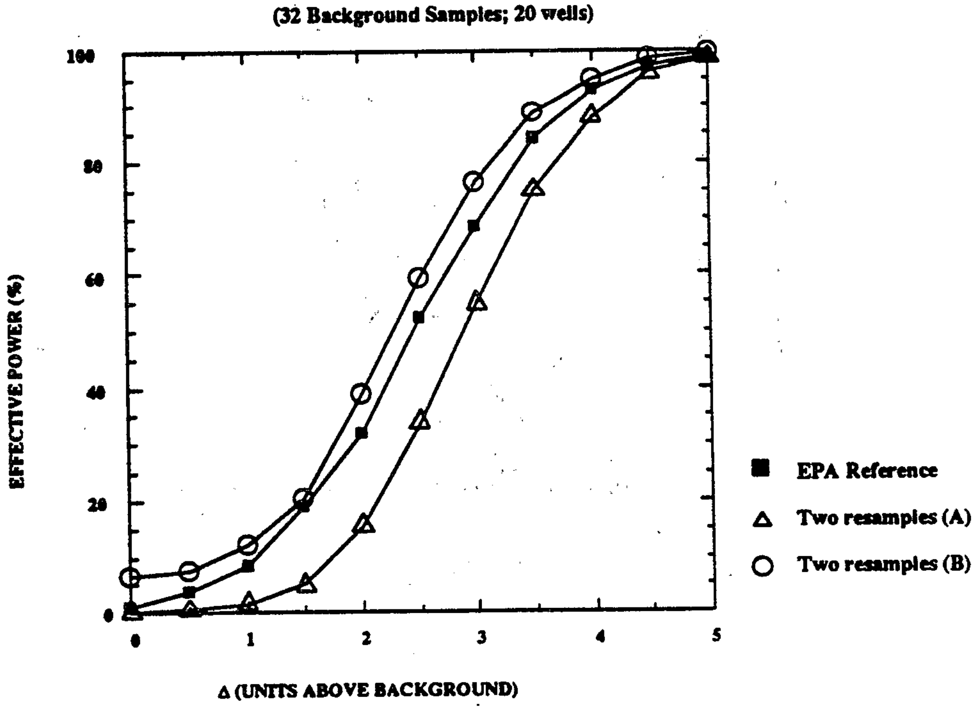


(24 Background Samples; 20 wells)

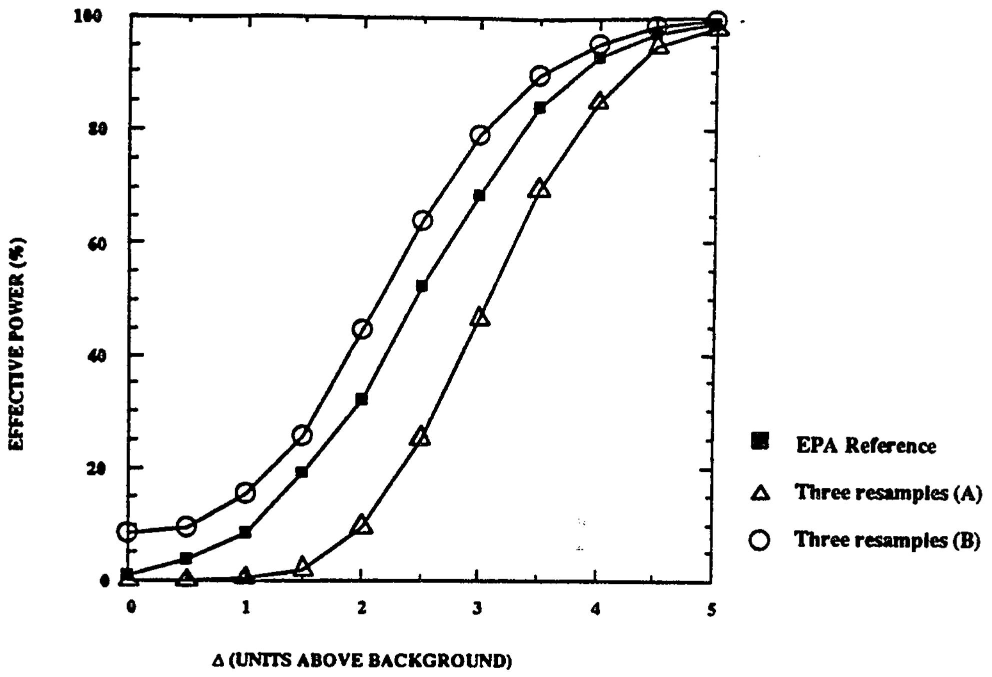


POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS



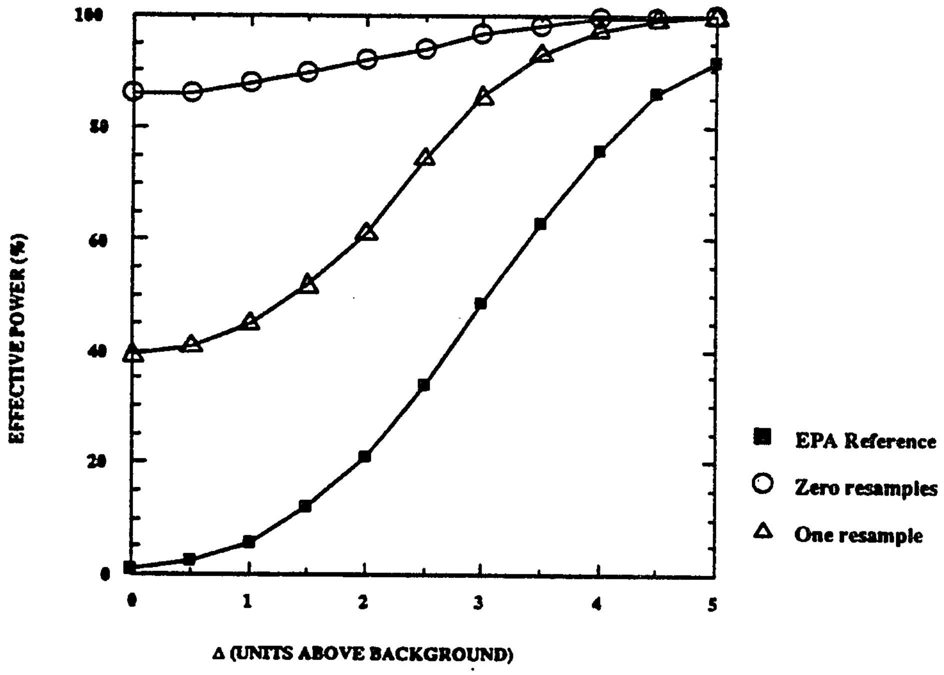


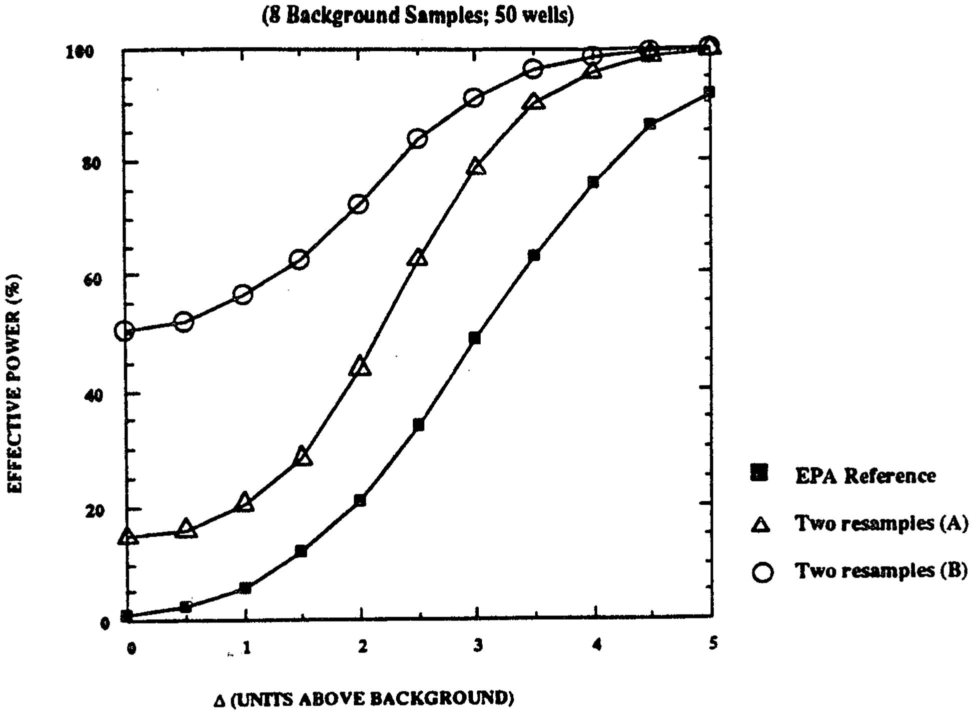
(32 Background Samples; 20 wells)



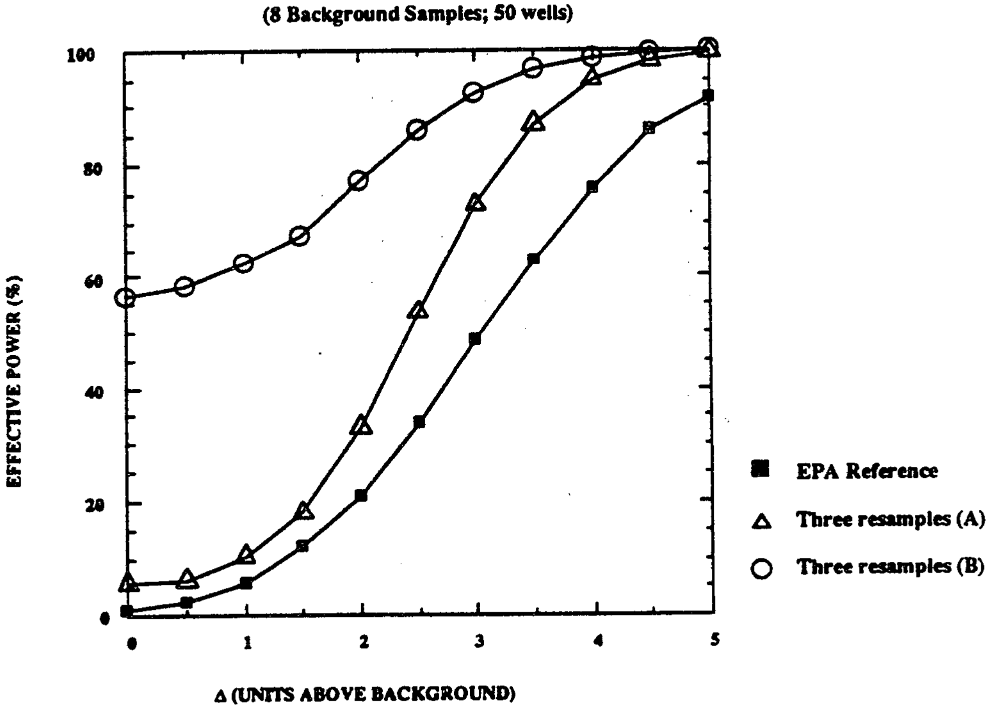
POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

(8 Background Samples; 50 wells)





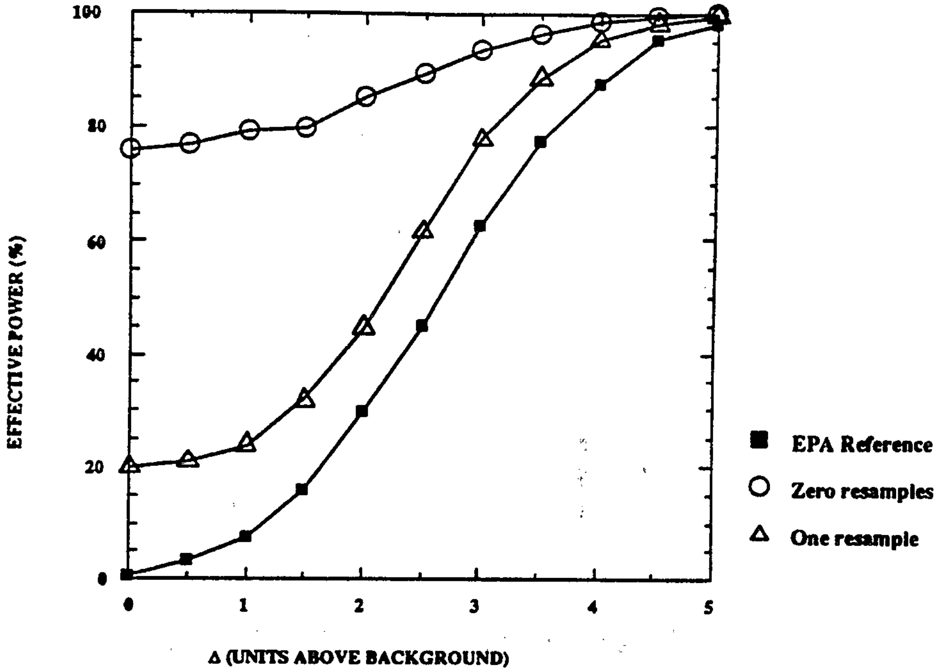
(8 Background Samples; 50 wells)
EFFECTIVE POWER (%)
0
20
40
60
80
100
0
1
2
3
4
5
Δ (UNITS ABOVE BACKGROUND)
■ EPA Reference
△ Two resamples (A)
○ Two resamples (B)



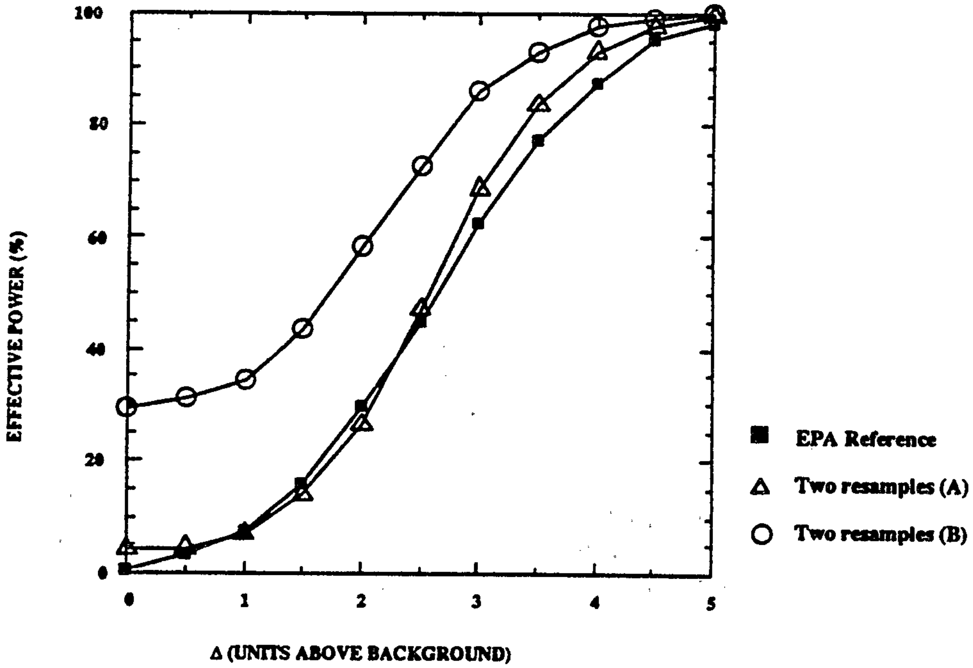
(8 Background Samples; 50 wells)
EFFECTIVE POWER (%)
0
20
40
60
80
100
0
1
2
3
4
5
Δ (UNITS ABOVE BACKGROUND)
■ EPA Reference
△ Three resamples (A)
○ Three resamples (B)

POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

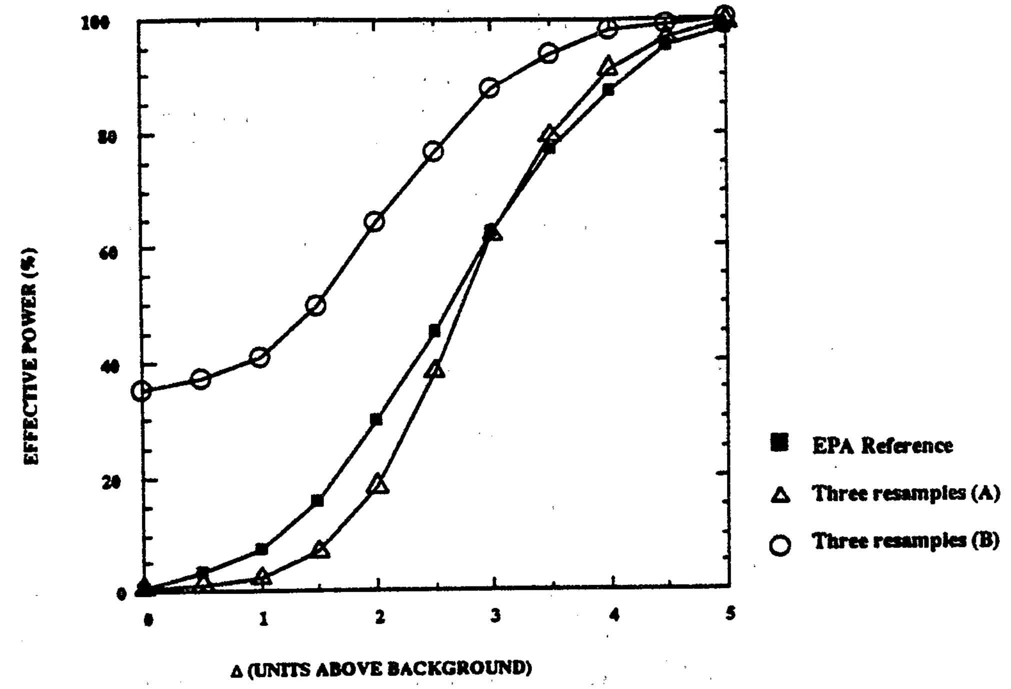
(16 Background Samples; 50 wells)



(16 Background Samples; 50 wells)

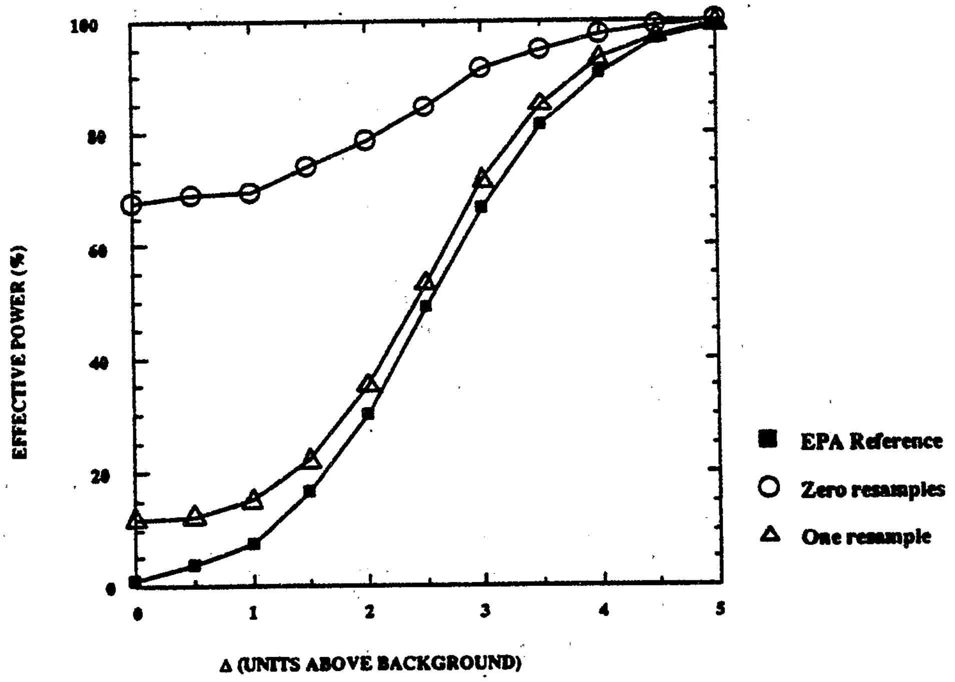


(16 Background Samples; 50 wells)

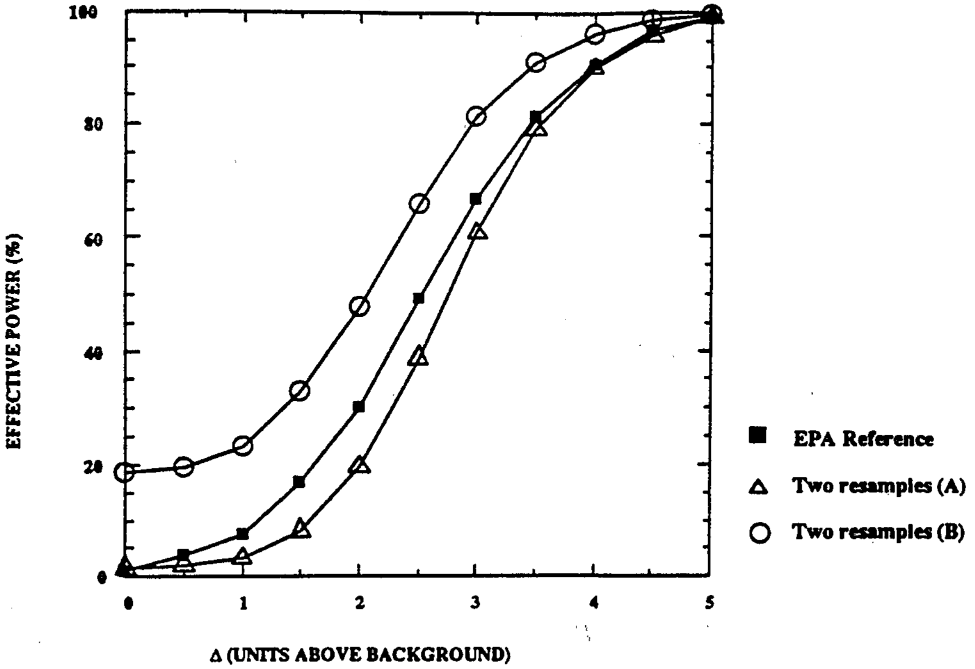


POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

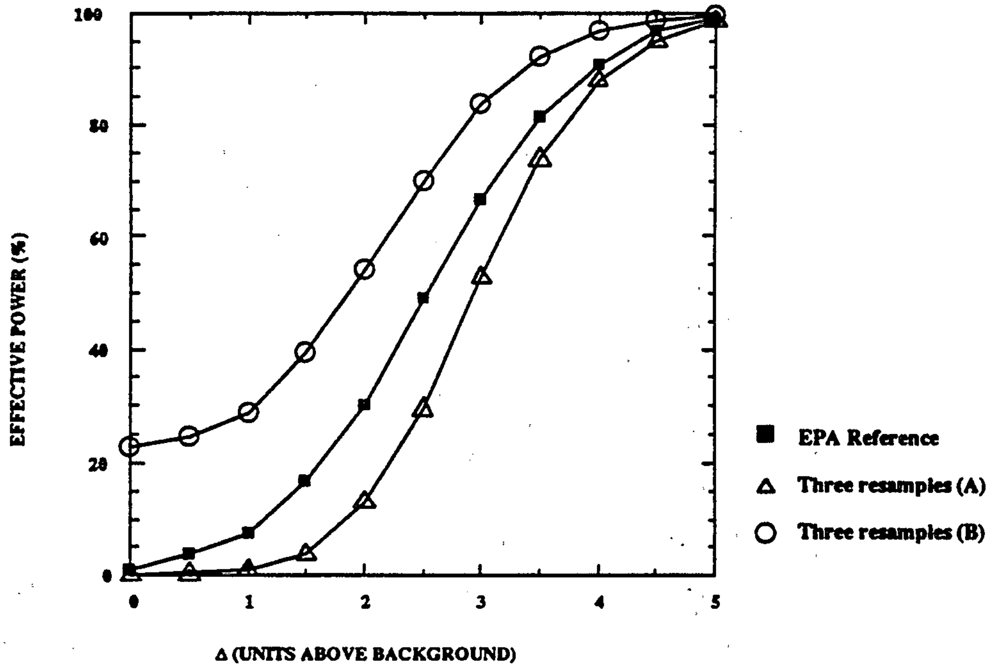
(24 Background Samples; 50 wells)



(24 Background Samples; 50 wells)

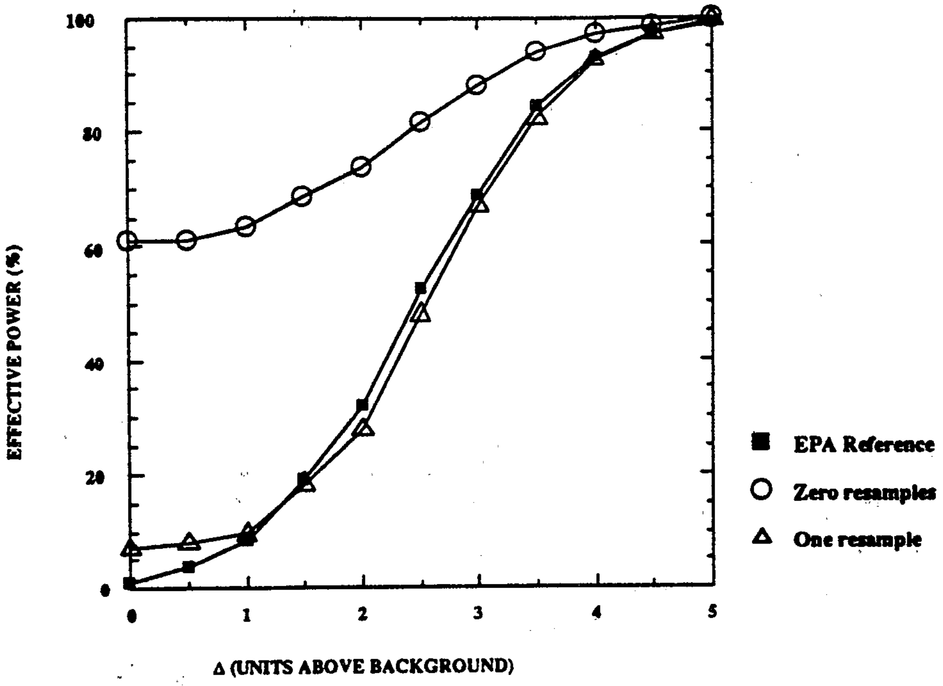


(24 Background Samples; 50 wells)

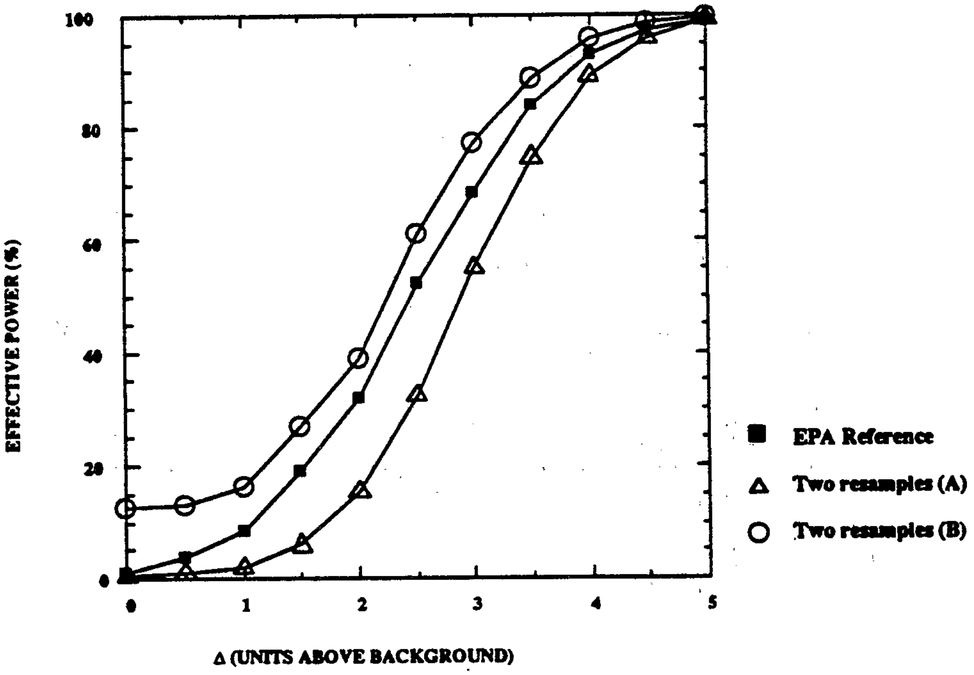


POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

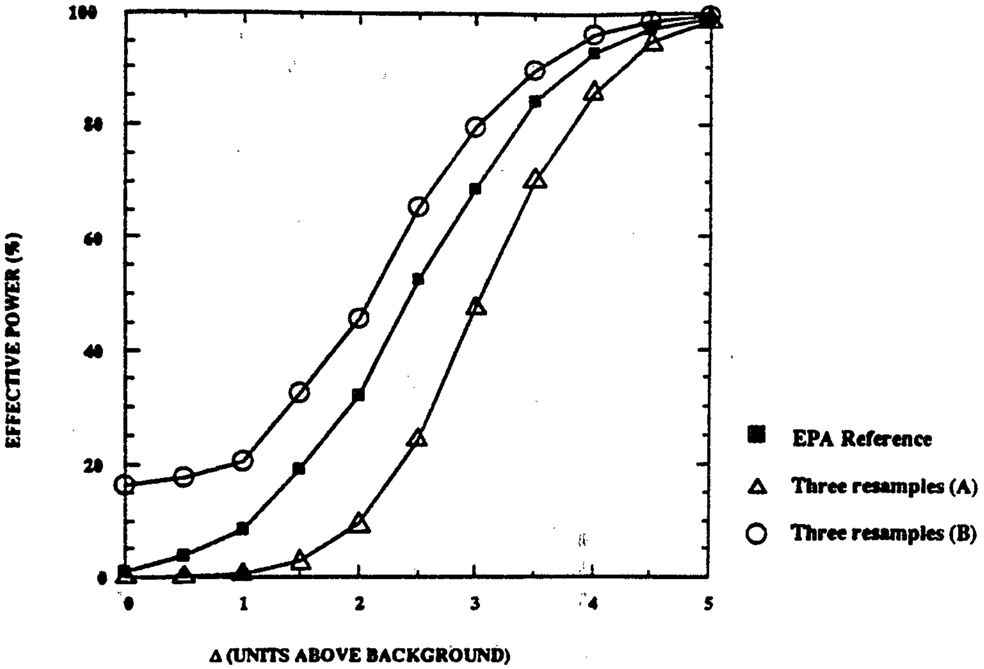
(32 Background Samples; 50 wells)



(32 Background Samples; 50 wells)

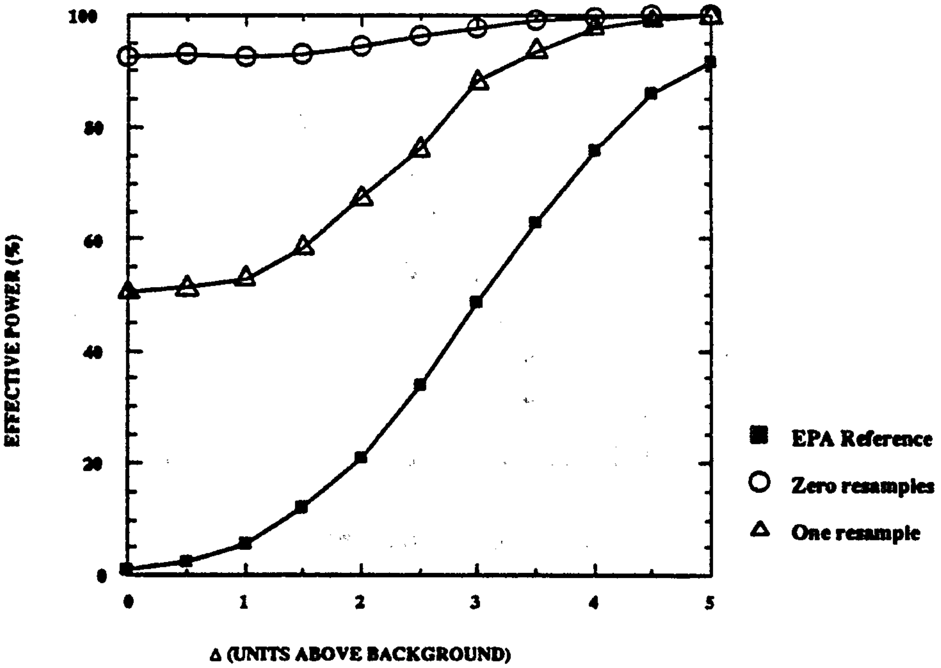


(32 Background Samples; 50 wells)

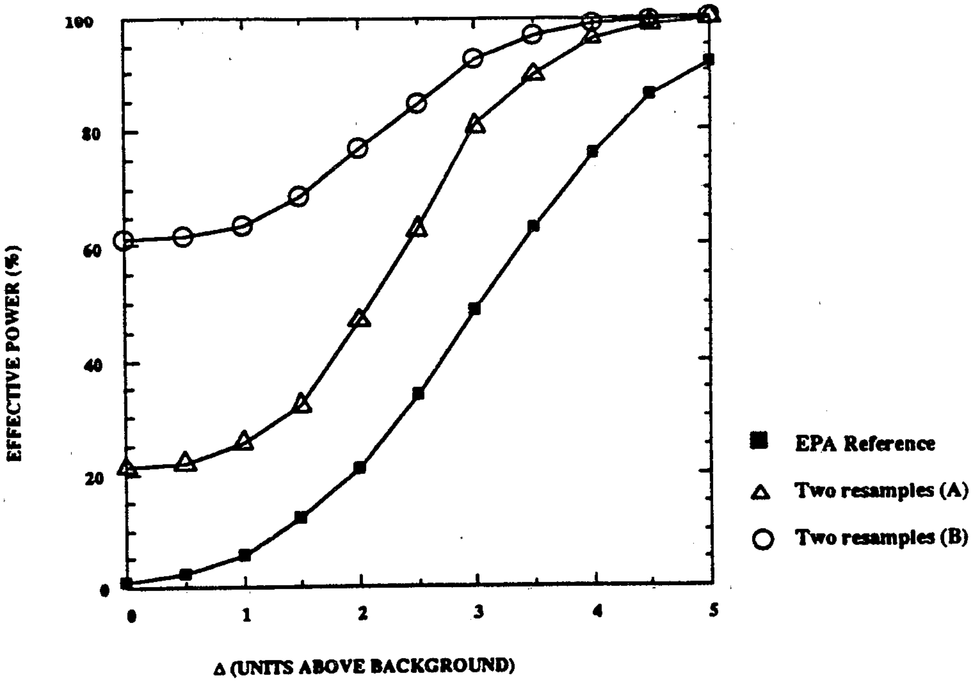


POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

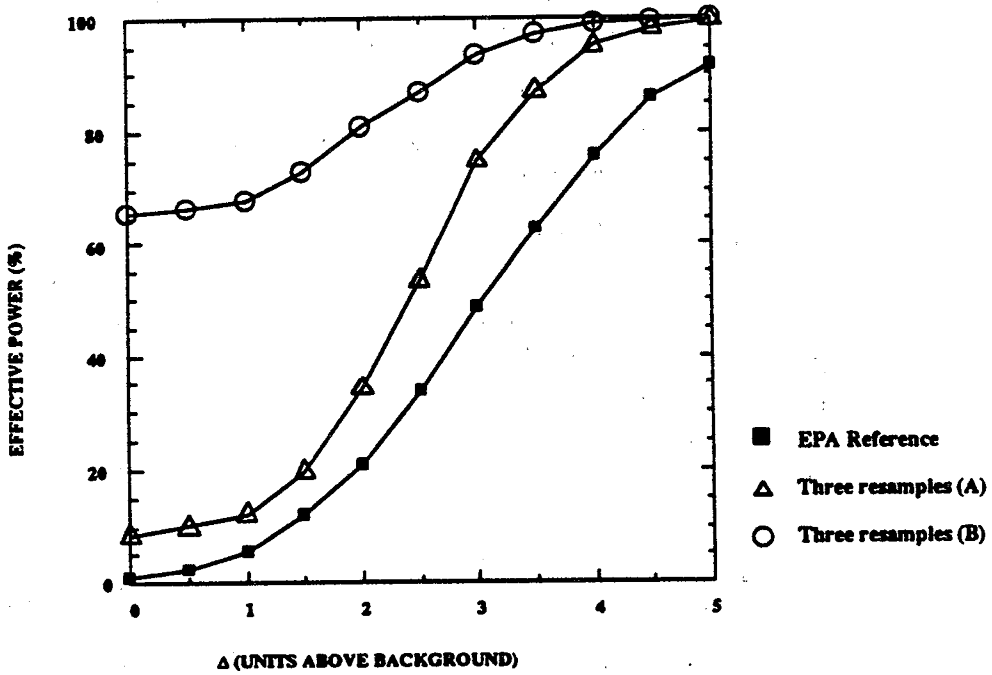
(8 Background Samples; 100 wells)



(8 Background Samples; 100 wells)

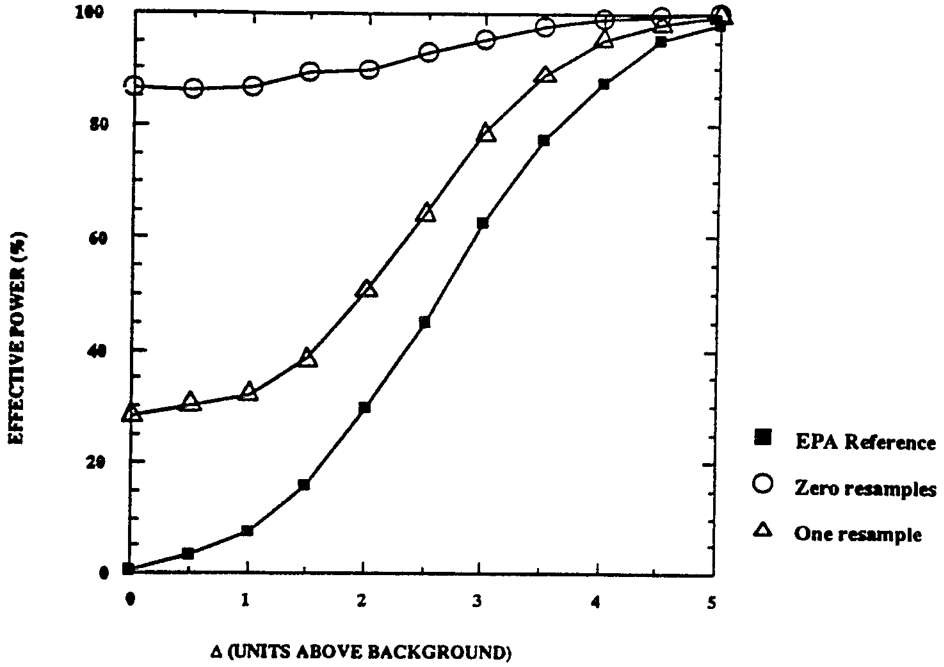


(8 Background Samples; 100 wells)

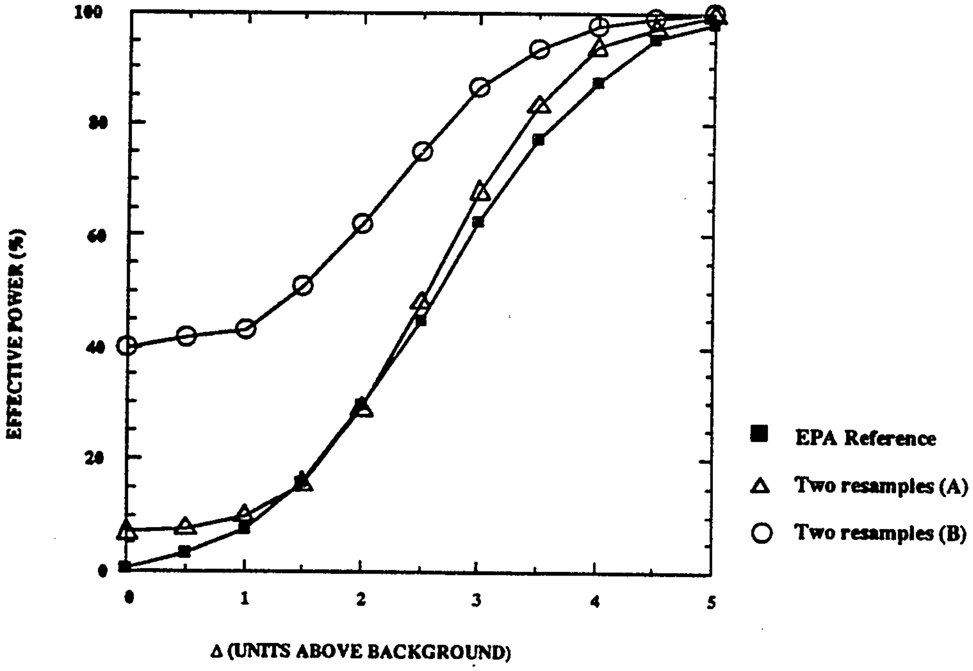


POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

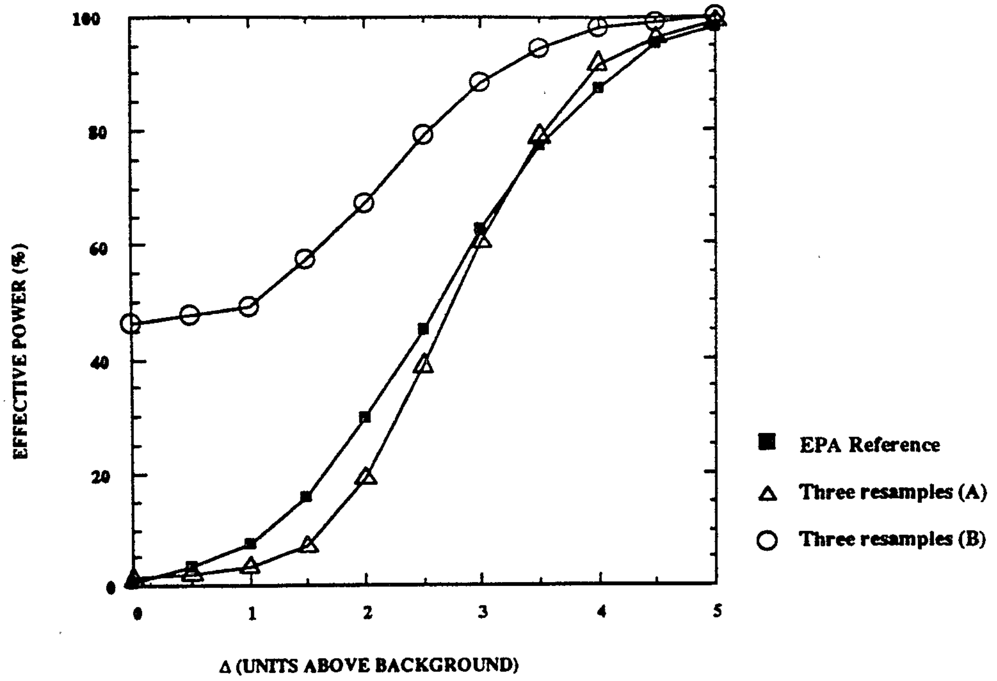
(16 Background Samples; 100 wells)



(16 Background Samples; 100 wells)

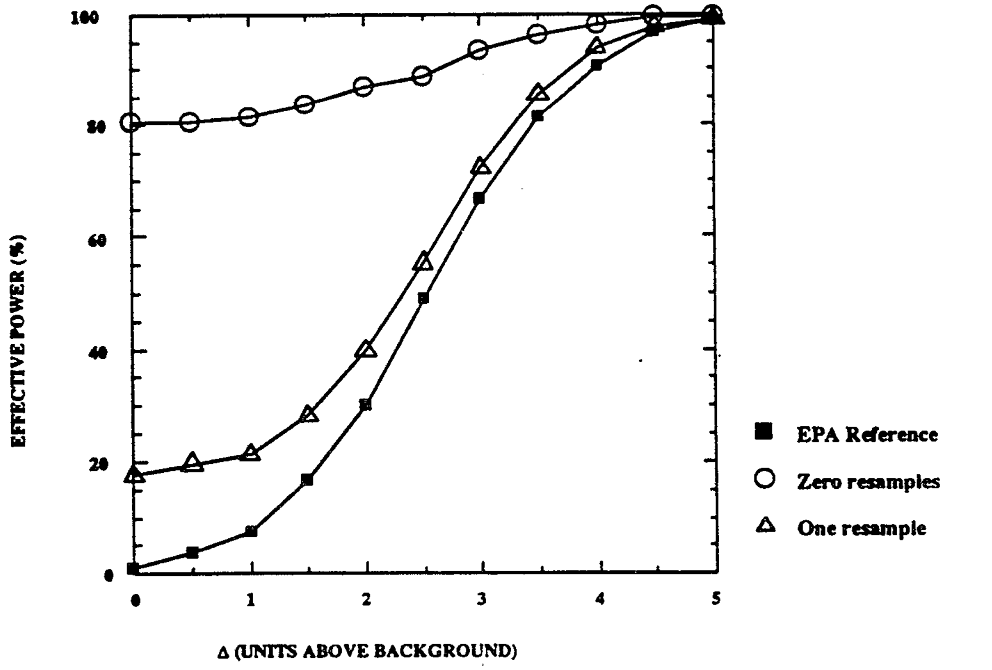


(16 Background Samples; 100 wells)

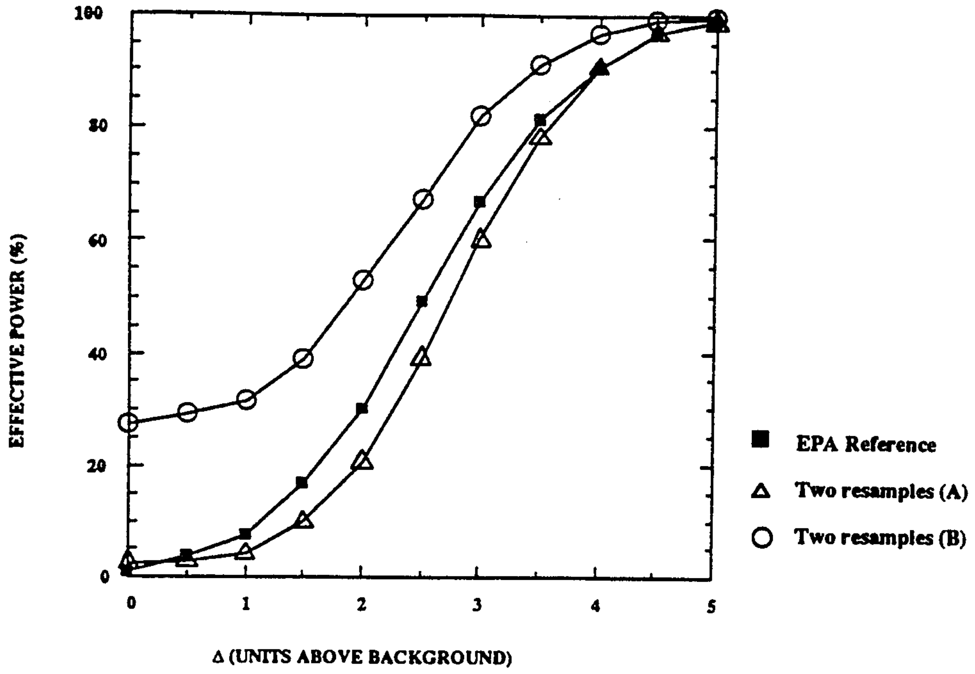


POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

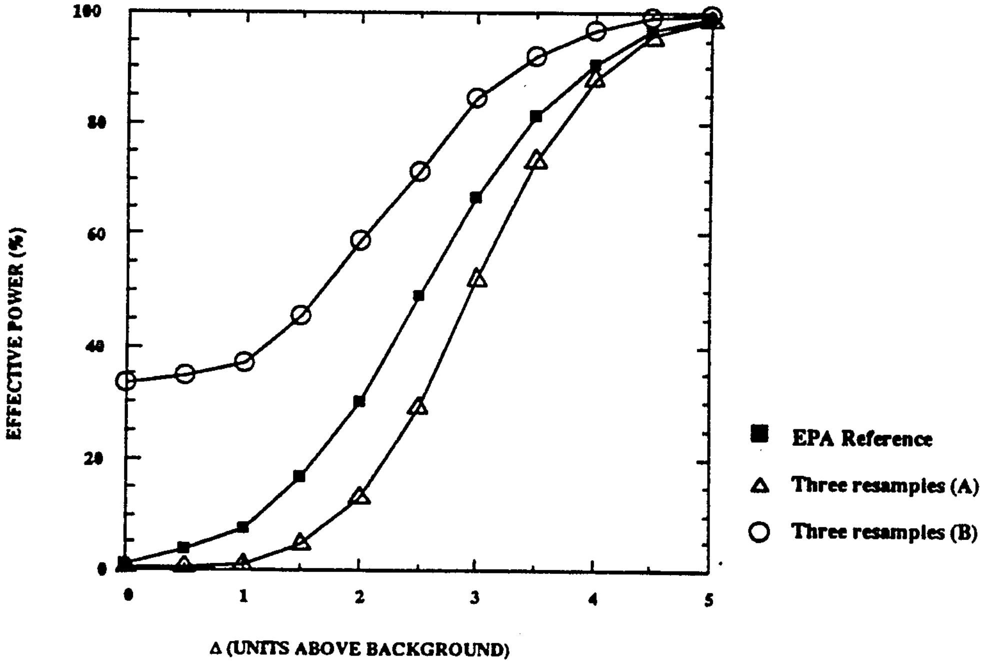
(24 Background Samples; 100 wells)



(24 Background Samples; 100 wells)

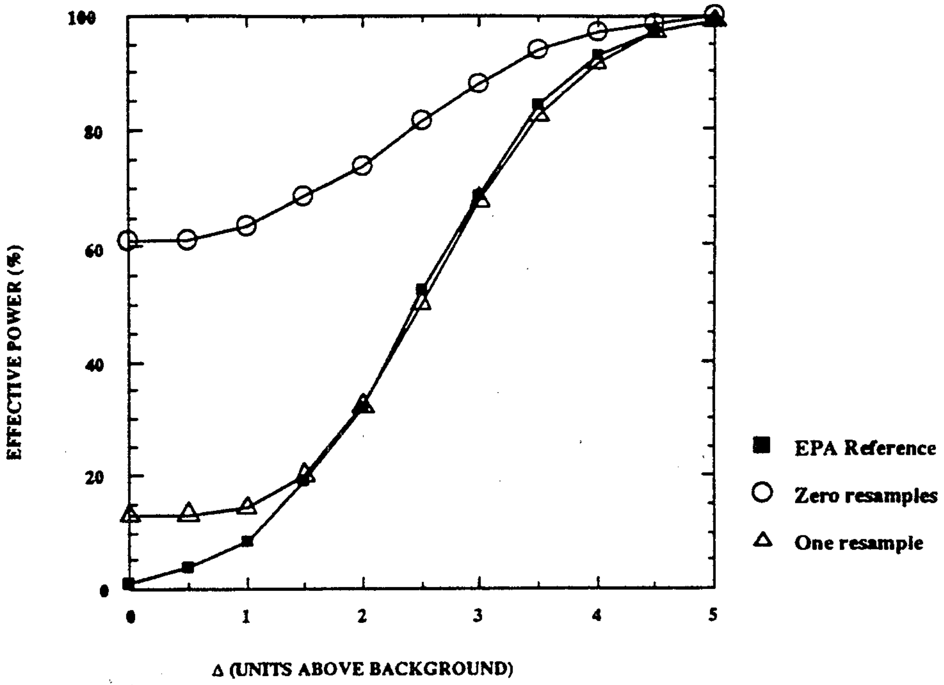


(24 Background Samples; 100 wells)

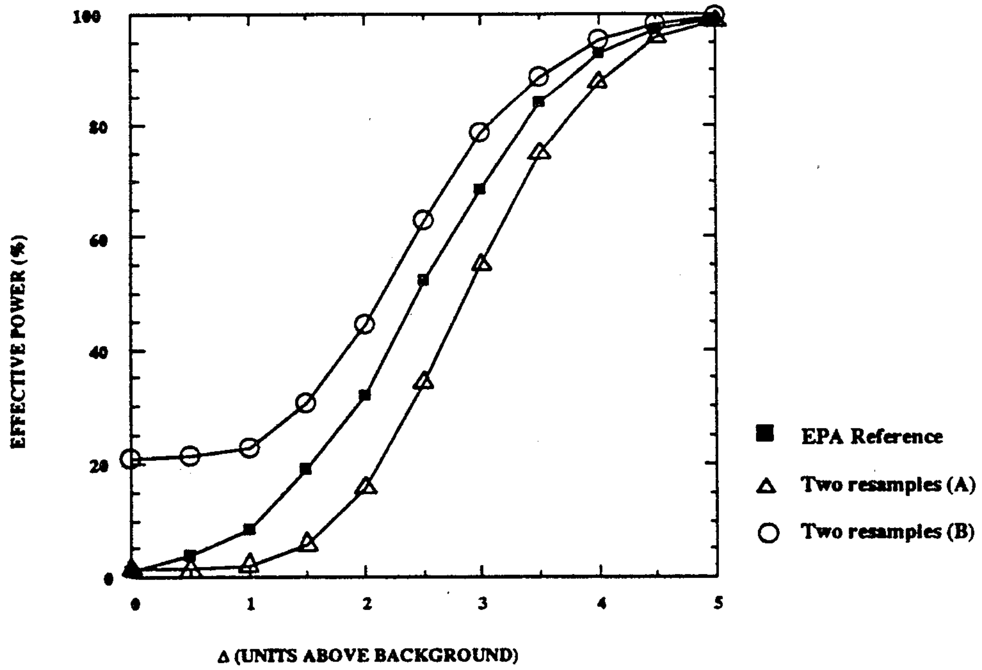


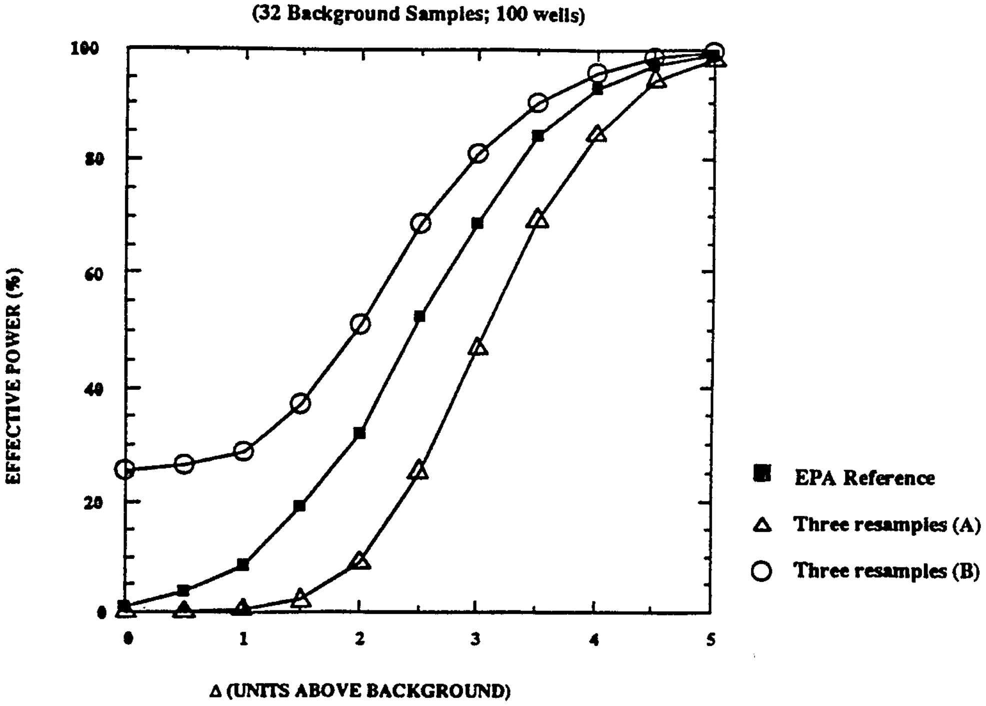
POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS

(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
Δ (UNITS ABOVE BACKGROUND)
■ EPA Reference
△ Three resamples (A)
○ Three resamples (B)

Section 1

"THE ONLY TROUBLE WITH A SURE THING IS UNCERTAINTY"

–Author Unknown

Excerpt from News and Numbers, A guide to Reporting Statistical Claims and Controversies in Health and Other Fields, by Victor Cohn. Iowa State University Press, Ames, Iowa, Copyright 1989, p.8.

EPA WORKSHOP SERIES

on the

Statistical Analysis of Groundwater Monitoring Data

featuring **GRITS/STAT**

WHAT YOU SHOULD LEARN

- Layman's guide to statistical regulations
- Introduction to GRITS/STAT software
- Intuitive understanding of statistical thinking and analysis
  - Expand your statistical vocabulary
  - Learn right questions to ask about your data
  - Apply basic distributional models

WHAT YOU SHOULD LEARN

- Basic techniques for statistical testing
  - ANOVA, t-tests
  - Confidence, tolerance, and prediction intervals
  - Control charts
  - Parametric versus non-parametric procedures
- Applying these techniques to groundwater data
  - How to use; when appropriate
  - Warnings, limitations, assumptions

SUMMARY OF REGULATIONS

- Statistical analysis of ground-water monitoring data at RCRA facilities: final rules
  - Subtitle C regulation: (53 FR 39720; October 11, 1988)
  - (hazardous wastes)
  - Subtitle D regulation: (56 FR 50978; October 9, 1991)
  - (municipal solid wastes)
- Interim Final Technical Guidance Document: April 1989
- Addendum to Interim Final Guidance: July 1992

WHY REPLACE THE CABF STUDENT'S T-TEST?

- Replicate sampling procedure
  - Assumed static background
  - Samples not independent
  - False positive rate too high
- Assumes normal distribution of original data
  - Can't handle frequent "non-detects"
- Can't handle large number of comparisons
  - Led to high false positive rates

"NEW" STATISTICAL PROCEDURES

- Offer more flexibility
- Sampling procedures based on site hydrogeology
- Can accommodate:
  - Departures from normality
  - Unequal variances
  - Temporal and spatial variability
  - Nondetects

"NEW" PROCEDURES INCLUDE

- Parametric and non-parametric ANOVA
- Parametric and non-parametric t-tests
- Confidence, tolerance, and prediction intervals
- Control Charts
- Alternative procedures

STATISTICAL REQUIREMENTS

- Test each constituent and well
- Choose statistical method from list of options
- Must comply with performance standards
  - Appropriate distributional model
  - Minimum false positive rates
  - Protect human health and environment
- Compare background versus downgradient data

TEST EACH CONSTITUENT/WELL

- Rationale: Identify the specific culprit(s)
- Regulations prohibit "pooling" of constituents
  - Constituents may have very different distributions
  - Often must test large number of constituents
  - Pooling many constituents can mask contamination

POOLING OF WELLS

- Regulations prohibit inappropriate "pooling" of wells
  - Don't lump data and discard well IDs
  - Hard to identify which well is culprit
- Appropriate "pooling" of wells OK
  - Use omnibus tests that keep well IDs intact
  - ANOVA is a good example of this strategy
  - Link statistical results to individual wells

CHOOSE FROM LIST OF METHODS

- Rationale: Too many tests to allow arbitrary choice
  - Best to have a few standard procedures
  - Alternative methods can be petitioned
- Standard tests include
  - ANOVA
  - Control Charts
  - Statistical Intervals

ANALYSIS OF VARIANCE (ANOVA)

- Background vs. one or more downgradient wells
  - With one downgradient well, equivalent to t-test
- Two-step procedure:
  - First run overall test
  - If significant, must test individual wells
- Parametric test uses original measurements
  - Do average levels differ among wells?
  - Non-parametric version uses ranks, tests medians

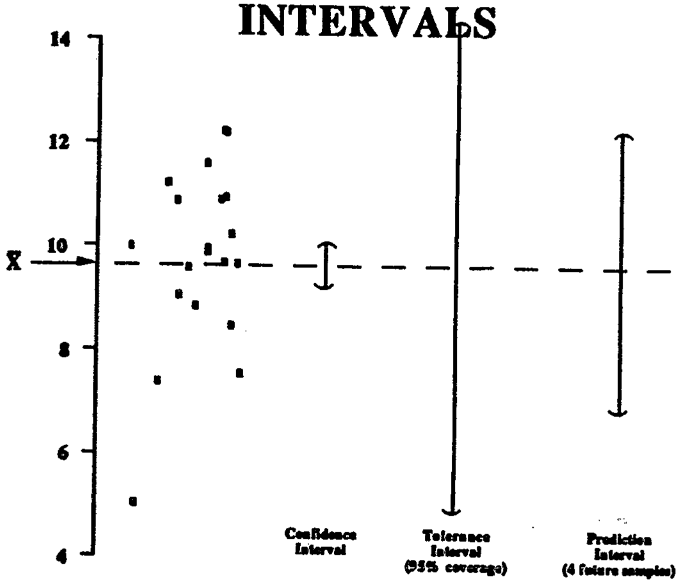
CONTROL CHARTS

- Single well plotted over time
  - Must be initially clean
- Visual tracking and identification of contamination
- Good for intrawell comparisons
  - e.g., in presence of spatial variability

STATISTICAL INTERVALS

- Three types: confidence, tolerance, and prediction
  - Different assumptions, different interpretations
  - Often used for special circumstances (e.g., retesting)
- When comparing against a regulatory standard (e.g., MCL)
  - use confidence intervals on mean or upper percentile
- Intrawell comparisons or limited compliance data
  - use prediction intervals or tolerance limits

STATISTICAL INTERVALS



PERFORMANCE STANDARDS

- Method must fit distribution of data
- Must meet minimum false positive rates
- Must be protective of human health and environment

MATCH TEST WITH DATA

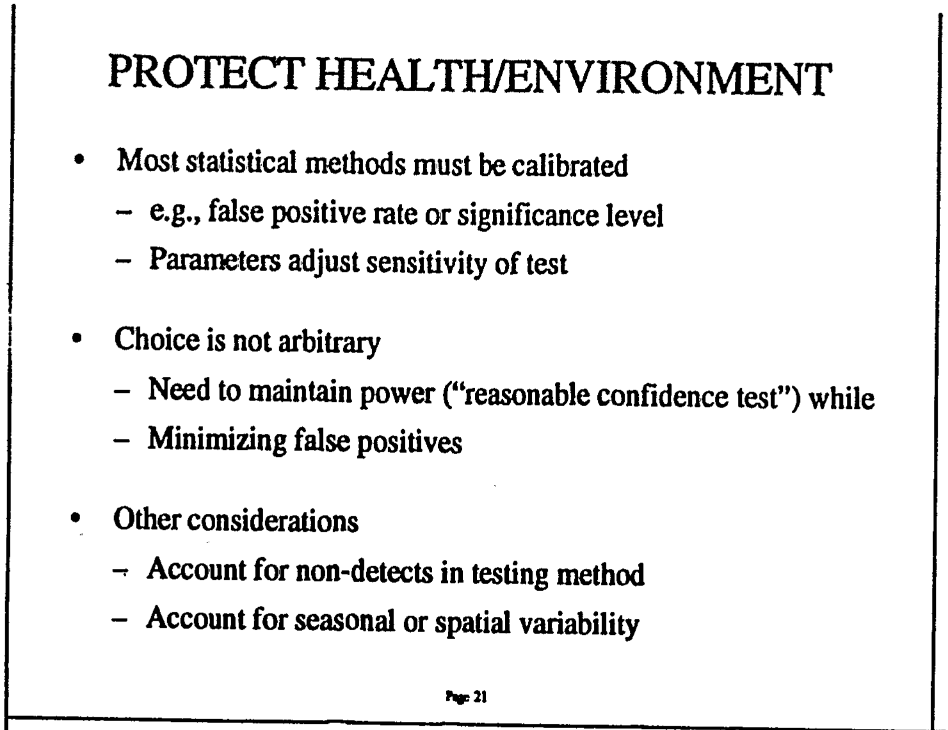
- All tests assume something about data distribution
  - Parametric: data normal or lognormal
  - Nonparametric: data symmetric; constant variance
- Meeting assumptions can be critical
  - Example: benzene data
- Transform data or change method if necessary
  - Normal versus lognormal data
  - Handling frequent non-detects

BENZENE DATA

Month	Background	Downgradient
1st	0.5 ppb	0.5 ppb
2nd	0.5 ppb	0.5 ppb
3rd	1.6 ppb	4.6 ppb
4th	1.8 ppb	2.0 ppb
5th	1.1 ppb	16.7 ppb
6th	16.1 ppb	12.5 ppb
7th	1.6 ppb	26.3 ppb
8th	0.5 ppb	186 ppb
mean	3.0 ppb	31.1 ppb

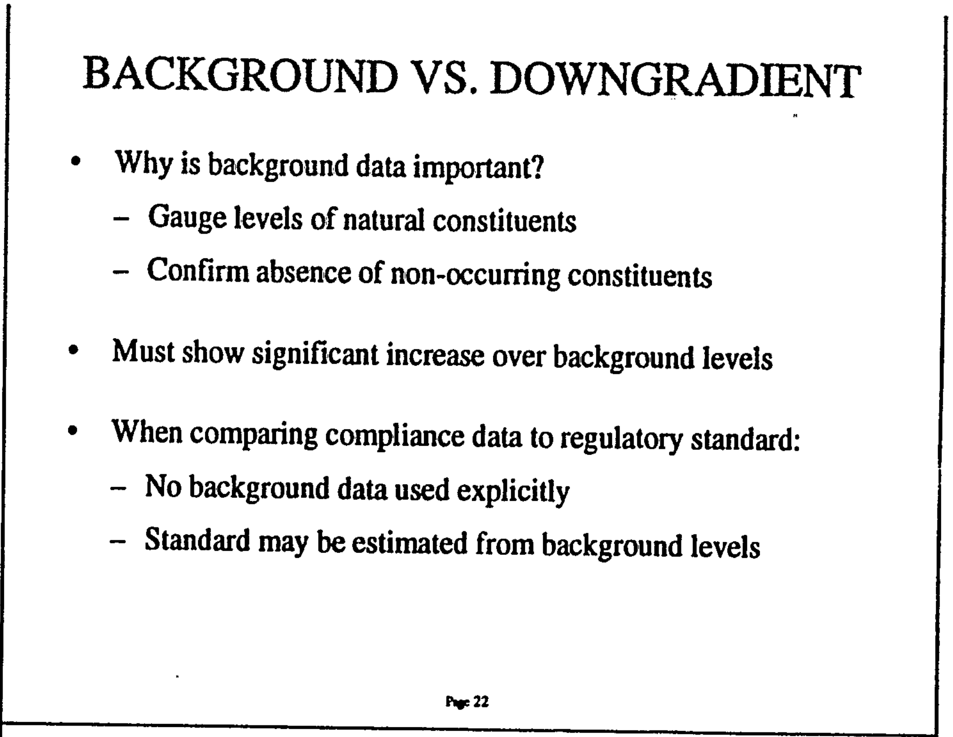
FALSE POSITIVE RATES

- Why is minimum rate necessary?
- Key: link between false positives, false negatives, and power
  - Statistical power: ability to identify real contamination
  - Power inversely related to false negatives
  - But, lower false positives also linked to lower power
- To maintain power, need minimum false positive rate
  - Statistical power is primary EPA concern



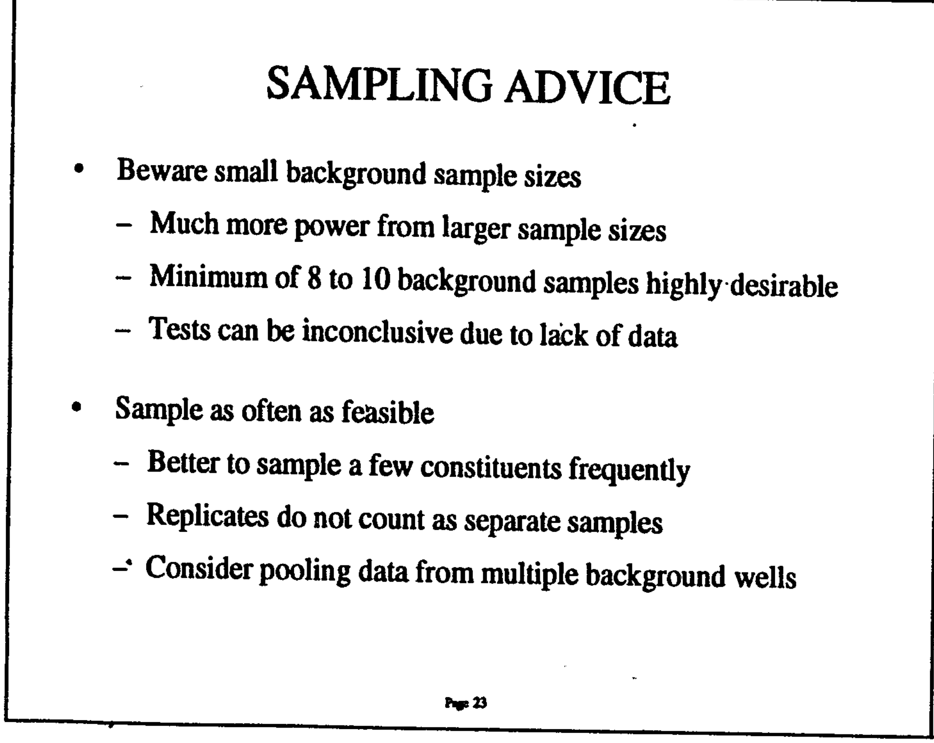
PROTECT HEALTH/ENVIRONMENT

- Most statistical methods must be calibrated
  - e.g., false positive rate or significance level
  - Parameters adjust sensitivity of test
- Choice is not arbitrary
  - Need to maintain power ("reasonable confidence test") while
  - Minimizing false positives
- Other considerations
  - Account for non-detects in testing method
  - Account for seasonal or spatial variability



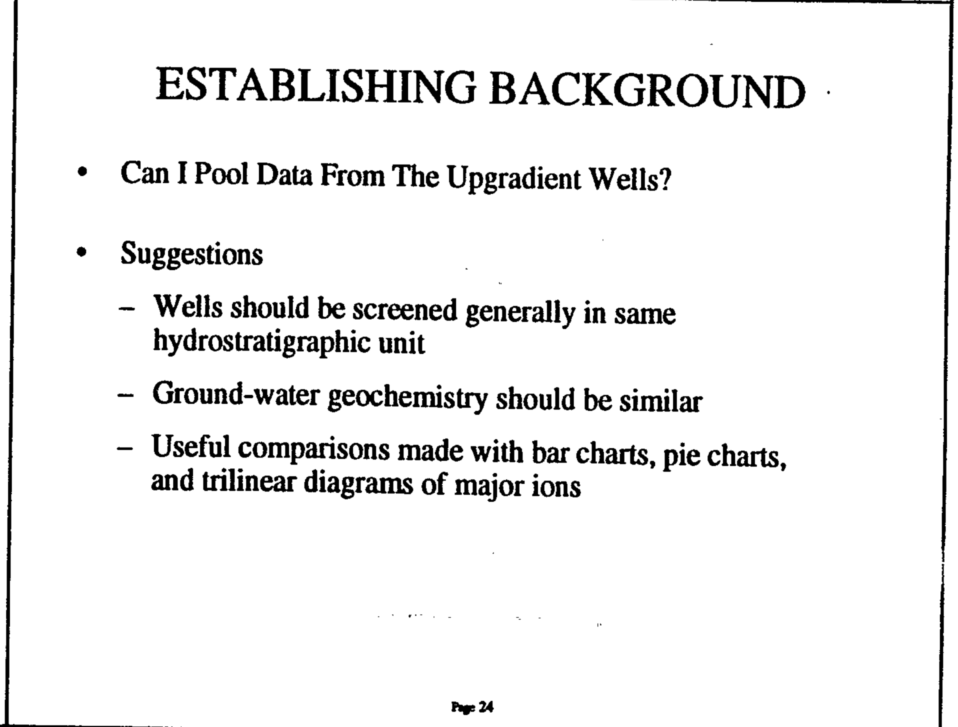
BACKGROUND VS. DOWNGRADIENT

- Why is background data important?
  - Gauge levels of natural constituents
  - Confirm absence of non-occurring constituents
- Must show significant increase over background levels
- When comparing compliance data to regulatory standard:
  - No background data used explicitly
  - Standard may be estimated from background levels



SAMPLING ADVICE

- Beware small background sample sizes
  - Much more power from larger sample sizes
  - Minimum of 8 to 10 background samples highly desirable
  - Tests can be inconclusive due to lack of data
- Sample as often as feasible
  - Better to sample a few constituents frequently
  - Replicates do not count as separate samples
  - Consider pooling data from multiple background wells



ESTABLISHING BACKGROUND

- Can I Pool Data From The Upgradient Wells?
- Suggestions
  - Wells should be screened generally in same hydrostratigraphic unit
  - Ground-water geochemistry should be similar
  - Useful comparisons made with bar charts, pie charts, and trilinear diagrams of major ions

Study and Interpretation of the Chemical Characteristics of Natural Water

Third Edition

By JOHN D. HEM

U.S. GEOLOGICAL SURVEY WATER-SUPPLY 2254

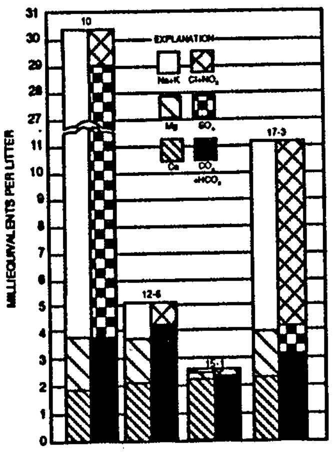


Figure 25. Analyses represented by bar lengths in milli-equivalents per liter. Numbers above bars indicate source of data in tables 10, 12, 15, and 17 (e.g., "12-6" — table 12, analysis 6).

(After Hem, 1989)

Hem, John D. 1989. Study and interpretation of the chemical characteristics of natural waters. United States Goelogical Survey Water-Supply Paper 2254. Third Edition. 263 pp.

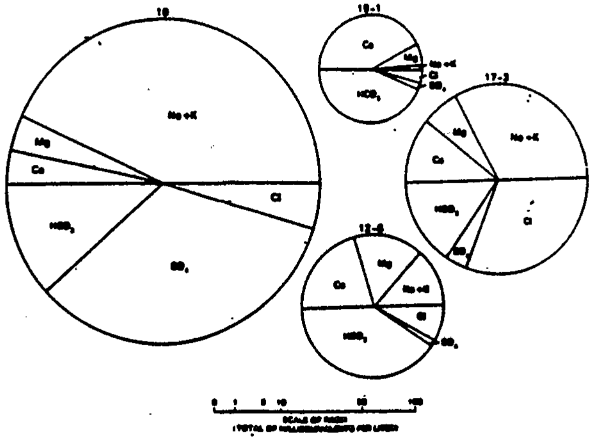


Figure 26. Analyses represented by circles subdivided on the basis of percentage of total milliequivalents per liter. Numbers above circles indicate source of data in tables 10, 12, 15, and 17 (e.g., "12-6"=table 12, analysis 6).

(After Hem, 1989)

Hem, John D. 1989. Study and interpretation of the chemical characteristics of natural waters. United States Geological Survey Water-Supply Paper 2254. Third Edition. 263 pp.

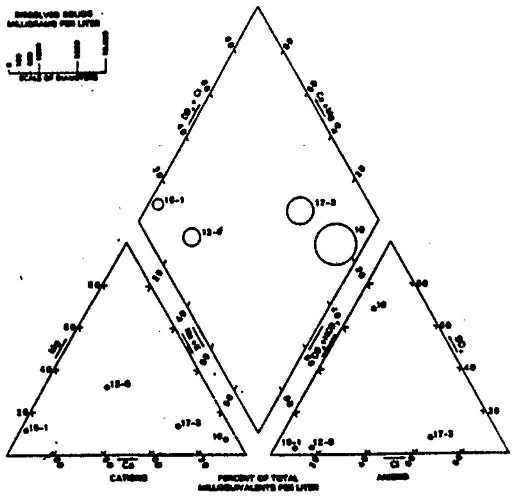


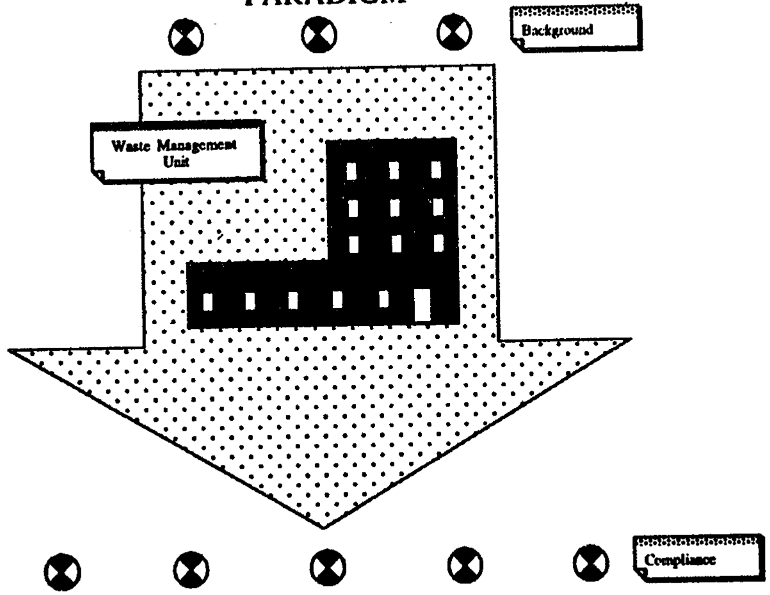
Figure 27. Trilinear diagram showing analyses represented by three-point plotting method. Numbers near circles indicate source of data in tables 10, 12, 15, and 17 (e.g., "12-6"=table 12, analysis 6).

(After Hem, 1989)

Hem, John D. 1989. Study and interpretation of the chemical characteristics of natural waters. United States Geological Survey Water-Supply Paper 2254. Third Edition. 263 pp.

Section 2

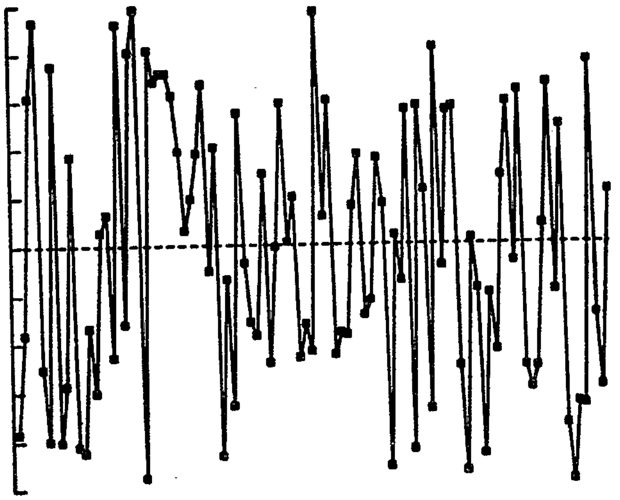
GROUNDWATER MONITORING PARADIGM



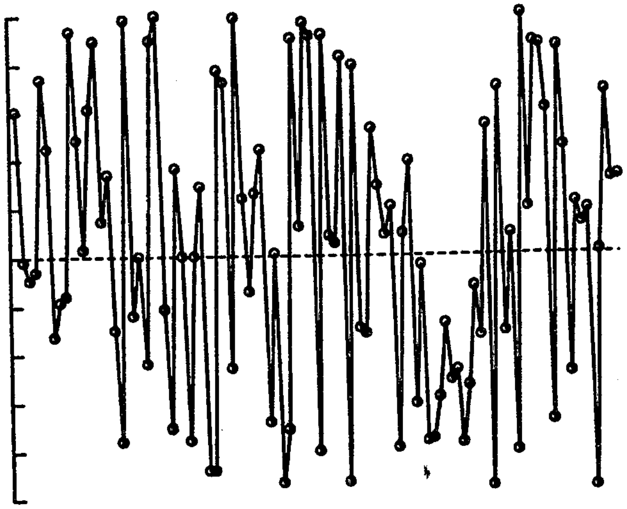
SOURCES OF VARIATION

- Random fluctuations
  - Sample variability from field handling/collection
  - Differences in repeated lab measurements
  - Natural variation in background levels
- Changes due to contaminant plume
- Need to separate random fluctuations from changes induced by contamination

BACKGROUND AVERAGES



COMPLIANCE AVERAGES



HYPOTHESIS TESTING

- To look for evidence of contamination, set up hypothesis test
  - H_0: No contamination at compliance well (**Null Hypothesis)**
  - H_A: Contamination has occurred (**Alternative Hypothesis)**
- Limited data usually available because of analysis costs
  - Creates statistical uncertainty
  - Decide which alternative better supported by sample evidence
  - Make decision based on strength of data

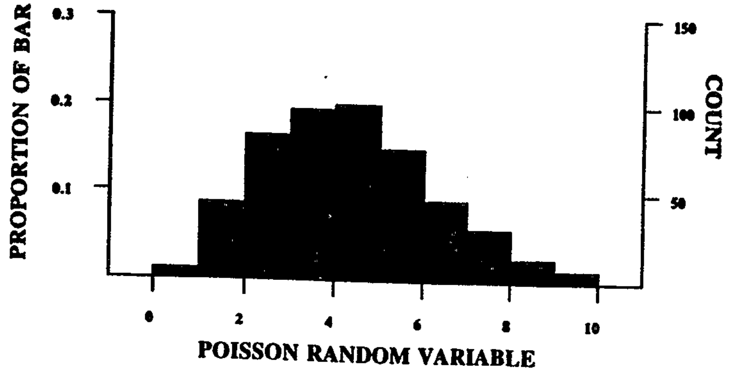
HYPOTHESIS TESTING

- Make sure hypothesis is matched to stage of monitoring
  - Detection monitoring hypothesis
    - H_0: No contamination
    - H_A: Contamination has occurred
  - Corrective action hypothesis
    - H_0: Contamination above action level
    - H_A: Contamination below action level

PROBABILITY DISTRIBUTIONS

- To make hypothesis tests work
  - Describe mathematical behavior of sample data
  - i.e., fit data to a probability distribution
- Probability distributions model random behavior
  - Approximation of reality
  - Can't predict specific results, but can determine how likely a given result is

POISSON DISTRIBUTION



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

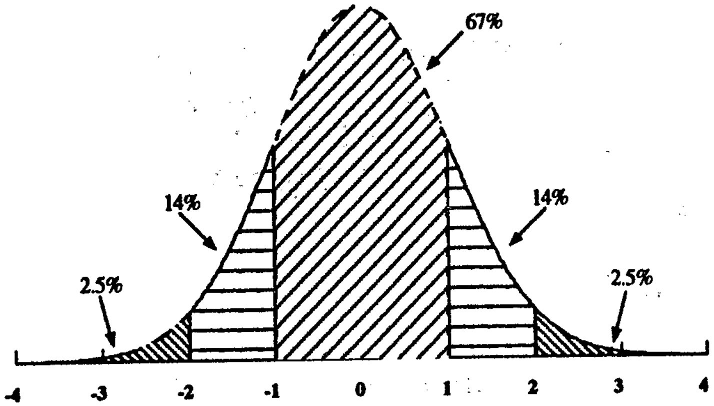
NORMAL DISTRIBUTION

- Data only called normal when they follow a specific equation

$$Pr\{X = x\} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- Under the normal curve:
  - Two-thirds of all values fall within 1 SD of the average
  - About 95% of all values fall within 2 SDs of the average
  - Only 5% of the values fall in the extreme tails
- The name "normal" does not mean other data are "abnormal"

STANDARD NORMAL DISTRIBUTION N(0,1)



CENTRAL LIMIT THEOREM

- Sums and averages of random variables tend to be normal
  - Even if original variables are <u>not</u> normal
- Examples: Body weights, spring loaded scale
- Because of CLT, normal distribution useful in testing
  - Often use arithmetic average to estimate population average
  - Need to know how the average will behave statistically

EXPLORATORY DATA ANALYSIS

- First step: explore data to find potential probability models
  - EDA consists of basic statistical tools/techniques
- Begin with time plots
  - Graphical method to view data at one or more wells
  - Can see trends and changes over time

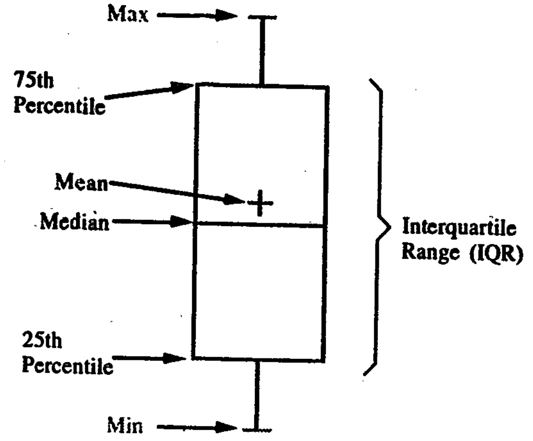
CENTER AND DISPERSION

- Need estimates of average behavior and variability to fit most probability distributions
- Numerical estimates of center
  - Mean and median
  - If different, data may be skewed or contain outliers
- Numerical estimates of dispersion
  - SD and interquartile range (IQR)
  - IQR = (75th – 25th percentiles)

BOXPLOTS

- Quick sketch of data distribution at one or more wells
  - Shows 25th, 50th, and 75th percentiles
  - GRITS/STAT version also gives min, max, and mean
- Range between ends of box equals IQR
- Can compare center and variability for multiple wells on same boxplot

BOX-WHISKER PLOT



Max

75th Percentile

Mean

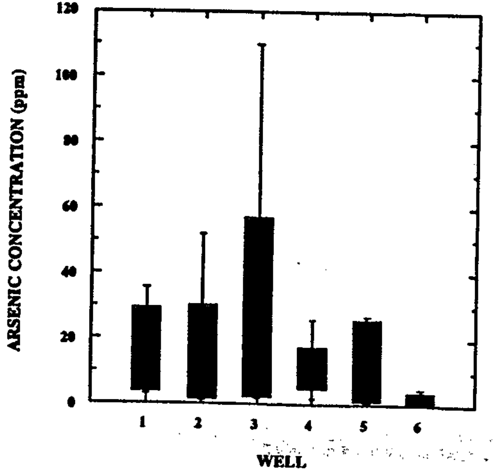
Median

25th Percentile

Min

Interquartile Range (IQR)

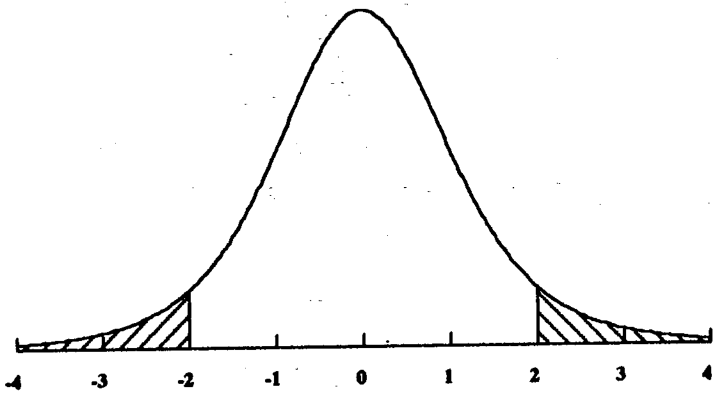
BOX PLOTS OF WELL DATA



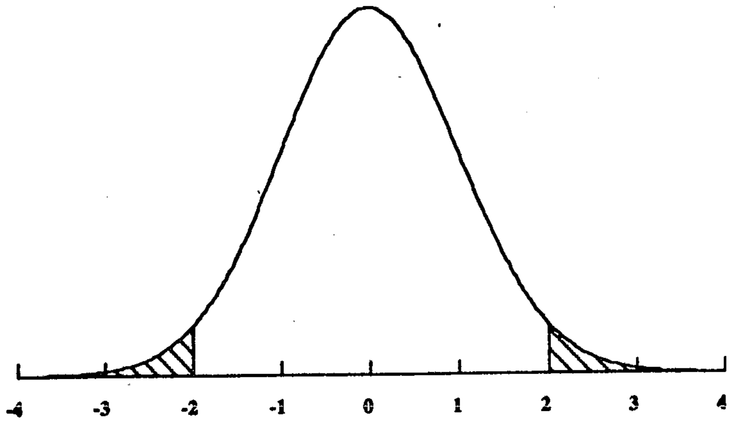
SYMMETRY VS. SKEWNESS

- Many types of data are symmetric (e.g., normal data)
  - But symmetry does not implу normality (e.g., t-distribution)
- Other data are skewed; must be fit to skewed distribution
  - Lopsided histogram; unbalanced tails
- Lognormal is a common positively-skewed distribution
  - Income patterns
  - Water quality concentration data
  - Key point: Logs of lognormal data are normal

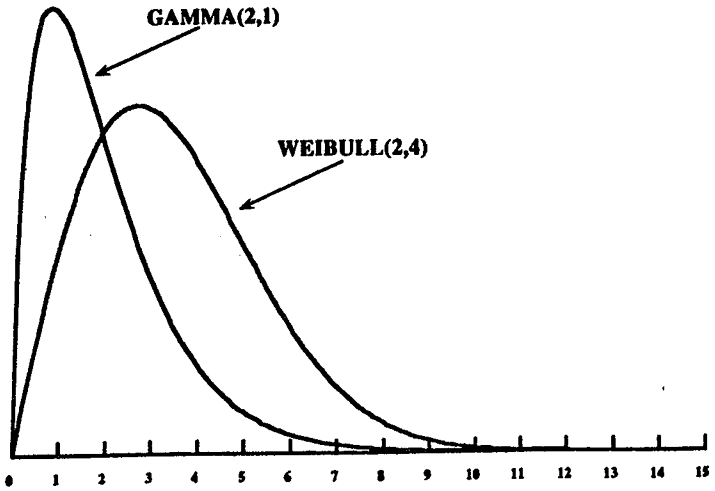
STUDENT'S t-DISTRIBUTION



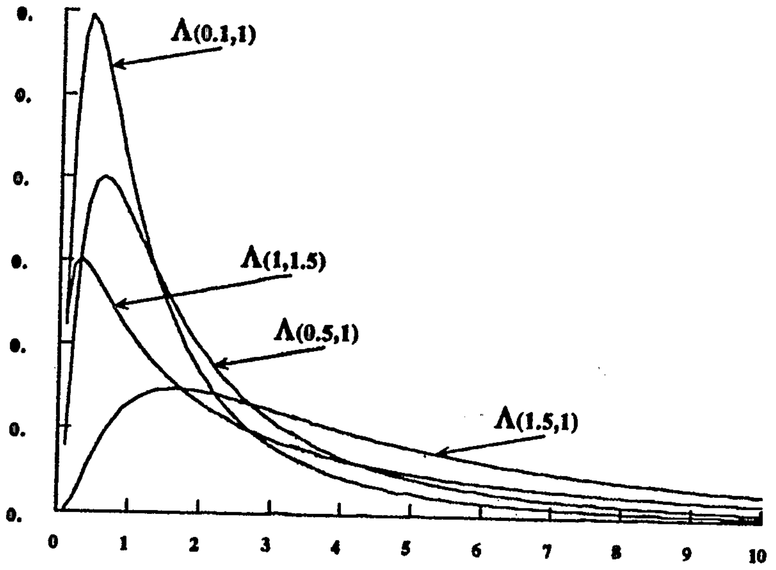
STANDARD NORMAL DISTRIBUTION



SKEWED DISTRIBUTIONS

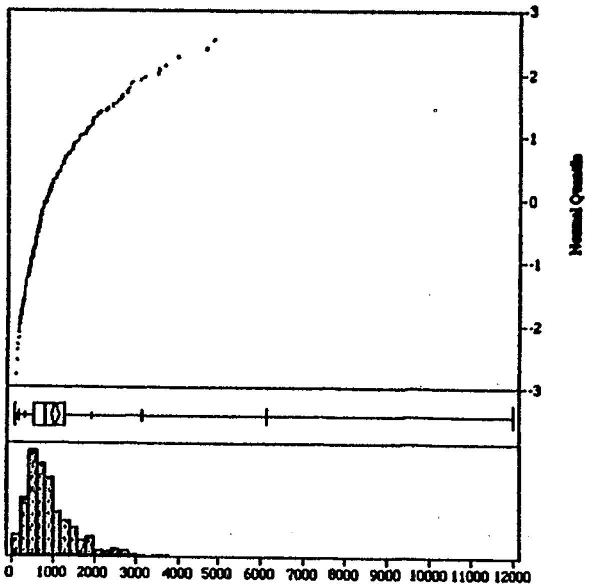


LOGNORMAL DISTRIBUTIONS



Page 21

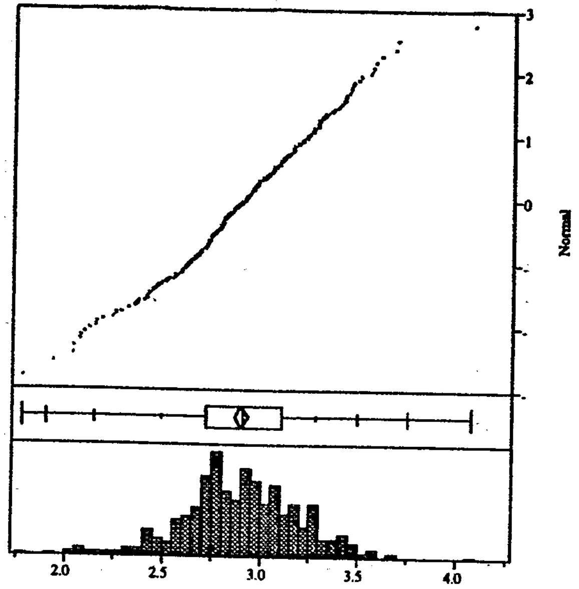
5-DAY BIOCHEMICAL OXYGEN DEMAND (BOD



Page 22

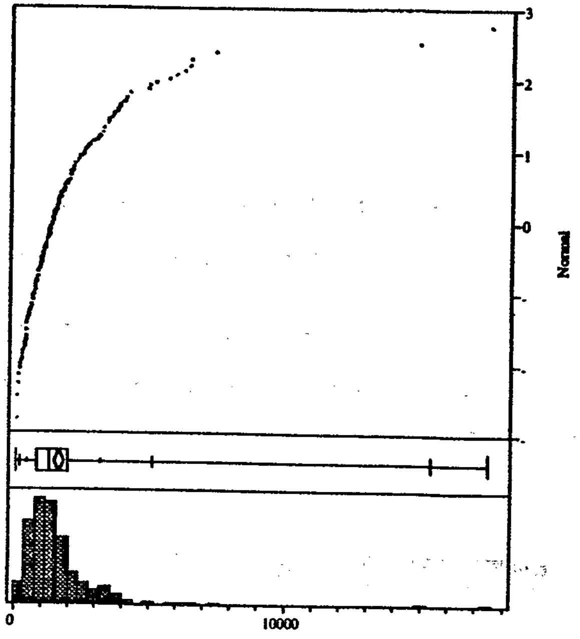
5-DAY BIOCHEMICAL OXYGEN DEMAND

Log Transformed



Page 23

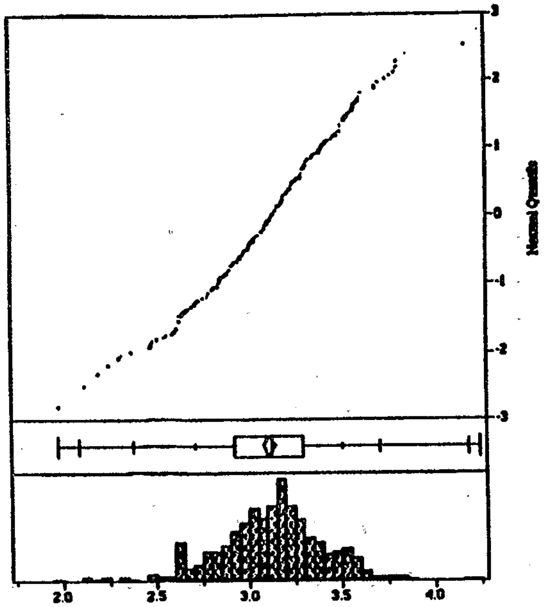
CHEMICAL OXYGEN DEMAND



Page 24

CHEMICAL OXYGEN DEMAND (COD)

Log Transformed



ESTIMATES OF SKEWNESS

- Skewness coefficient
  - Represents average cubed deviation from sample mean
  - Symmetric data will have skewness close to zero
  - Asymmetric data will have positive or negative skewness
- Highly skewed data indicate non-normal pattern
  - Lognormal data have positive skewness
- CV sometimes used to measure skewness, but not recommended
  - CV of logged values often unreliable

PROBABILITY PLOTS

- Visual comparison of data to a probability model
  - Often used to decide between normal and lognormal
- Plot of ordered sample values vs. normal z-scores
- Directly shows departures from normality
  - skewness, outliers, etc.
- Straight line fit indicates normal data
  - Linear fit of logs implies lognormal data

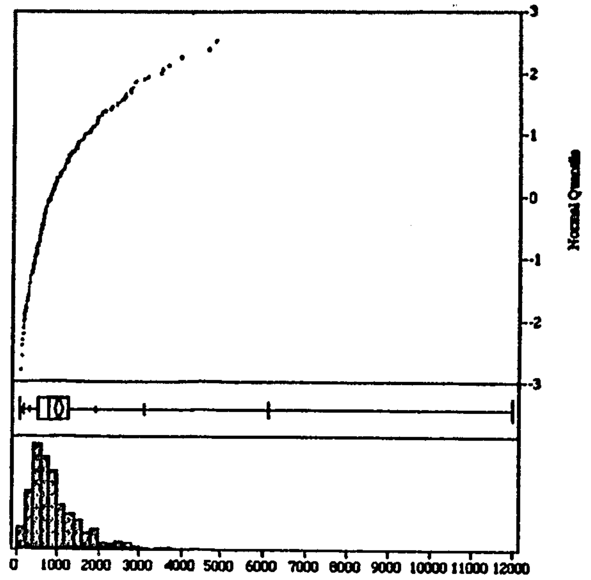
LEAD CONCENTRATION (ppb)

Quarter	BW 1	BW 2	BW 3	BW 4
1	2.5	10.7	7.9	7.6
2	6.6	6.7	12.4	21.0
3	13.5	10.4	6.8	7.2
4	27.0	7.5	7.7	3.7
5	9.9	23.0	5.2	5.9

Ordered LEAD Concentrations

Lead Concentration (ppb)	Order (I)	Cumulative Probability 100*(I/(n+1))	Normal Quantile
2.5	1	5	-1.669
3.7	2	10	-1.309
5.2	3	14	-1.068
5.9	4	19	-0.876
6.6	5	24	-0.712
6.7	6	29	-0.566
6.8	7	33	-0.431
7.2	8	38	-0.303
7.5	9	43	-0.18
7.6	10	48	-0.06
7.7	11	52	0.06
7.9	12	57	0.18
9.9	13	62	0.303
10.4	14	67	0.431
10.7	15	71	0.566
12.4	16	76	0.712
13.5	17	81	0.876
21.0	18	86	1.068
23.0	19	90	1.309
27.0	20	95	1.669

5-DAY BIOCHEMICAL OXYGEN DEMAND (BOD$_5$



Section 3

HYPOTHESIS TESTING BASICS

- Set up formal test between competing alternatives
- Which alternative best supported by data?
  - Are compliance data similar to background or not?
- Hypothesis testing similar to criminal trial
  - One hypothesis initially favored over the other
  - Initial hypothesis rejected only with strong evidence

STEPS INVOLVED

1. Set up H_0
   - The observed data or statistic will follow a known distribution
   - H_0 represents the assumed or favored condition
   - Example of H_0: concentration of suspected pollutant is zero
2. Set up H_A
   - Under H_A, the data or statistic will follow a distribution different from H_0
   - Example of H_A: concentration of suspected pollutant is large, leading to a measured concentration >0

STEPS INVOLVED (cont.)

3. Take measurements and calculate statistic(s)
4. Compare results with distribution predicted under H_0
   - If probability of observed result is very small (typically less than 5% or 1%) then either
     - An unlikely event occurred because of random variation,

       or
     - The null hypothesis, H_0, is incorrect

GAMBLER'S RUIN

Problem: Gambler wants you to bet on "Tails;" Won't let you examine coin but will flip it 10 times for free.

H_o: Coin is fair $\Rightarrow P(H) = \frac{1}{2} = P(T)$

H_A: Coin is biased; $P(H) > \frac{1}{2} \Rightarrow P(T) < \frac{1}{2}$

Set up Test: Reject H_o if Number of Heads is "too big"

$$P(\#H = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{if } P(H) = p$$

$$\left.\begin{aligned} P_o[10H] &= 0.0009766 \\ P_o[9H] &= 0.009766 \\ P_o[8H] &= 0.04394 \end{aligned}\right\} \alpha = 0.0547$$

$$P_o[7H] = 0.11719$$

TEST RESULTS

# Heads	
0	
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

Power = Probability of rejecting $\mathbf{H_O}$ when $\mathbf{H_O}$ is false
Depends on what $\mathbf{H_A}$ holds

Example : **P(H) =.8** Under $\mathbf{H_A}$:

$$\left.\begin{array}{l} \mathbf{P_A(H=10)=0.0107374} \\ \mathbf{P_A(H=9)=0.26843} \\ \mathbf{P_A(H=8)=0.30199} \end{array}\right\} 0.5778 \text{ Power}$$

WHAT CAN GO WRONG?

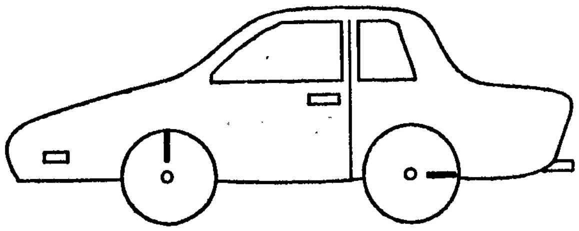
- As in legal trial, have two alternatives, two distinct errors
  - Can hang the innocent or free the guilty
- In statistical terms:
  - Accept H_A when H_0 is true (false positive)
  - Accept H_0 when H_A is true (false negative)

DECISION

TRUTH		**Accept H_0**	**Accept H_A**
	H_0	**OK**	**TYPE I ERROR**
	H_A	**TYPE II ERROR**	**OK**

ERRORS IN HYPOTHESIS TESTING

- Probability of a **Type I error, α**
  - Deciding contamination present when groundwater not contaminated
  - α = significance level of the test
- Probability of a **Type II error, β**
  - Failing to detect contamination when present
- Often work with complement of β
  - 1-β = **power of test**
- Power of test depends on
  - Significance level, α
  - Amount of data
  - How polluted the well is
  - Large concentration differences easier to detect than small ones



H_0: Innocent ⇒ Wheel marks lined up by chance

How likely is H_0?

Wheels in same position as marked

$$\text{Prob}(H_0) = \frac{1}{12} * \frac{1}{12} = \frac{1}{144}$$

What are type I and II errors in this case?

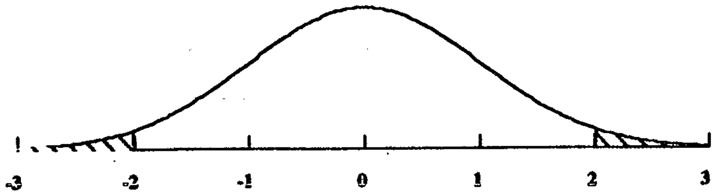
NUTS & BOLTS: SAMPLING DISTNS

- Key: find chance that result could have been seen under H_0
  - Need to determine distributional behavior of test statistic
  - Use sampling distributions to do this
- Statistics like mean or sum are random variables, too
  - Why? Built from individual random values
- Probability distribution for a test statistic called sampling distribution
  - May be different than distribution for individual values

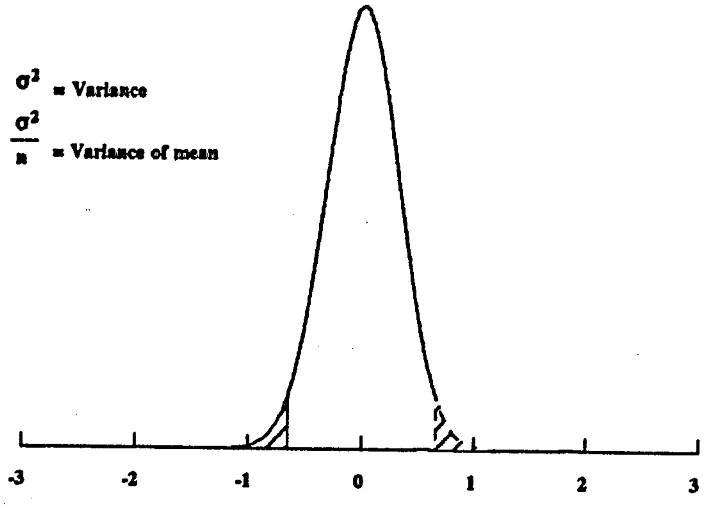
CLT AND SAMPLING DISTNS

- By CLT, sums and averages will have normal sampling distributions
  - Many test statistics will therefore follow normal pattern
- Not just any normal distribution, however:
  - Variance of mean depends on sample size
  - Variance of mean less by a factor of n
  - Can predict behavior of mean with greater accuracy

NORMAL DISTRIBUTIONS N(0,1)



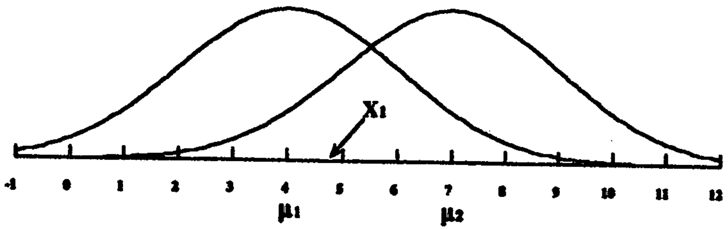
NORMAL DISTRIBUTION N(0,0.1)



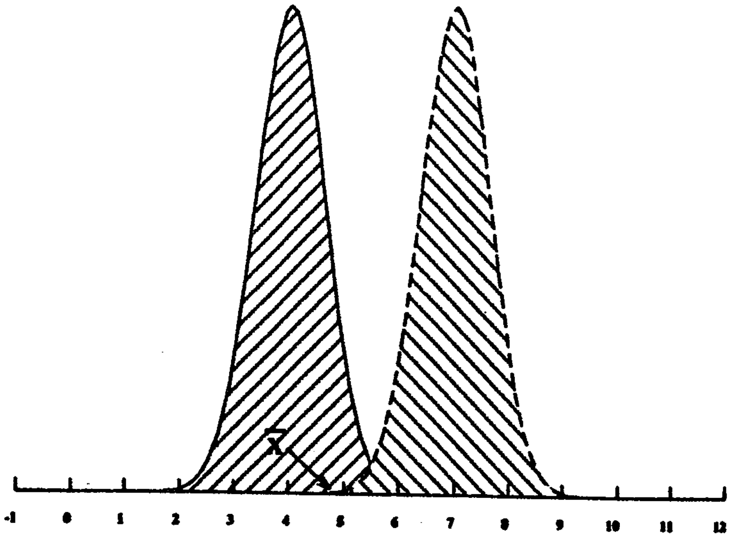
SAMPLING DISTN: SO WHAT?

- Need to squeeze information about underlying population from limited sample data
- Sampling distribution a big help because:
  - Variability of sample mean much less than variability of any single measurement, and
  - Distributional behavior of mean is known by CLT
- Can better pinpoint the location of true population mean
  - Standard error (SE) is really the SD of sampling distn

ORIGINAL DISTRIBUTIONS

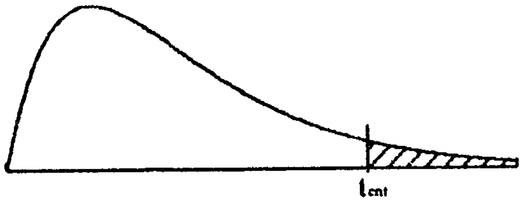


VS. SAMPLING DISTRIBUTIONS



WHY STUDY SAMPLING DISTNS?

- First Goal: Combine observations into a summary statistic
- Sampling distribution describes the behavior of this statistic under H_0
  - — Can gauge whether calculated value is too extreme
  - — Behavior under H_0 is critical
- Data $x_1, x_2, \ldots x_n$ Statistic: T_n



Distribution of T_n

WHY STUDY SAMPLING DISTNS?
(continued)

- If T_n is too extreme, H_0 will be rejected
- Sometimes T_n not a natural summary, but used because sampling distribution is known

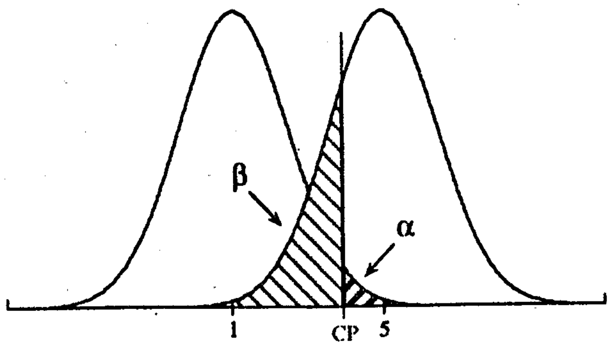
TYPE I ERROR: FALSE POSITIVES

- α = Probability of wrongly accepting H_A when H_0 is true
- Want to minimize false alarm rate
  - – Consider a smoke detector
- For a given sample size, α and β are linked
  - – t-test example
- Strategy
  - – Set α at say 5% or 1%; choose critical point using α

SAMPLING DISTNS OF MEAN

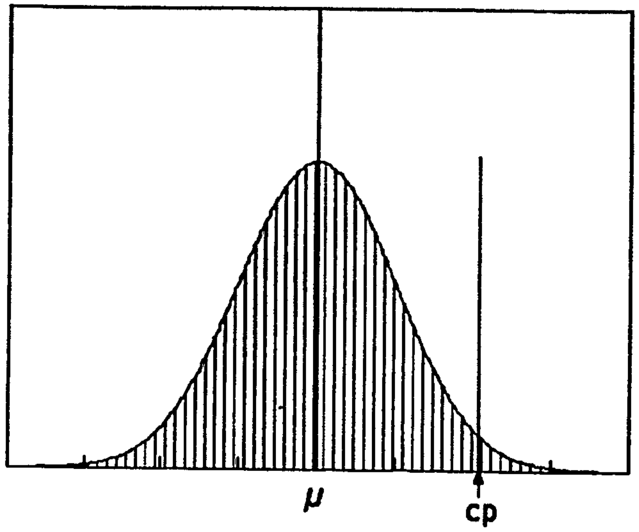
H_0 : μ = 1 ppb

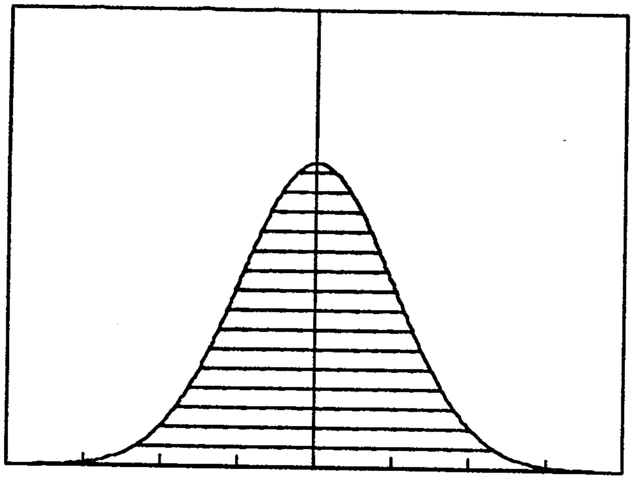
H_A : μ = 5 ppb

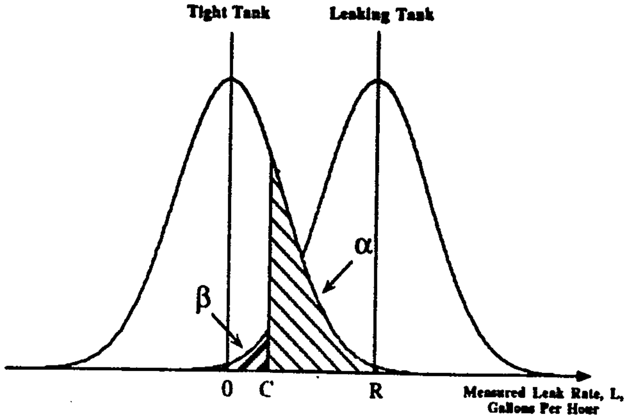


FALSE NEGATIVES AND POWER

- Type II error (β) is the chance of a false negative
  - False negatives of concern in AIDS blood testing
  - In groundwater, β = probability of missing true contamination
  - 1-β = power (sensitivity) of test to detect contamination
- For a fixed α, false negative rate depends on
  - Level of contamination (easier to detect large differences)
  - Sample size
- Minimizing false negative rate often more important than minimizing false positive rate







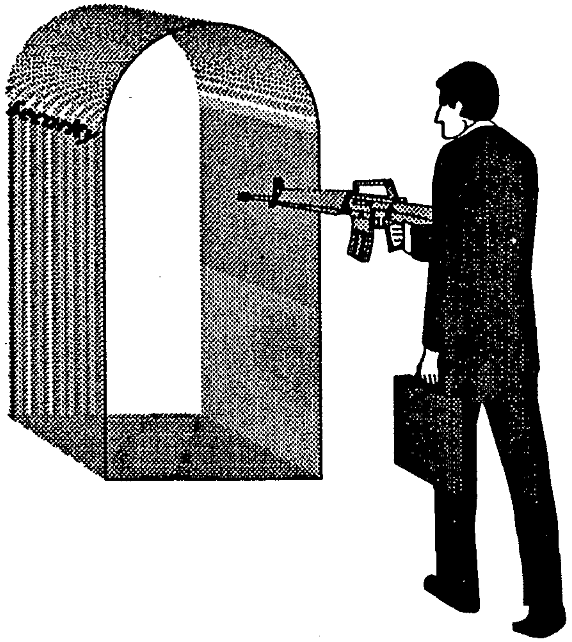
C = Criterion or threshold for declaring a leak (a leak is declared if the measured rate exceeds C)

α = Probability of False Alarm, P(FA)

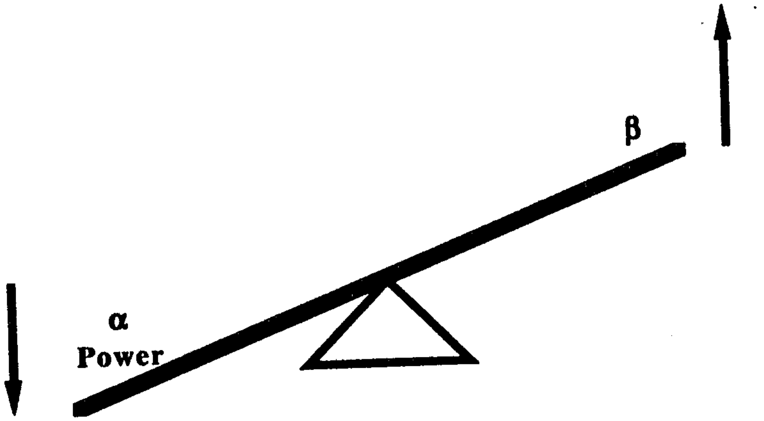
β = Probability of not detecting a leak rate R

1 -β = Probability of detecting a leak rate R, P(D(R))

R = Leak Rate



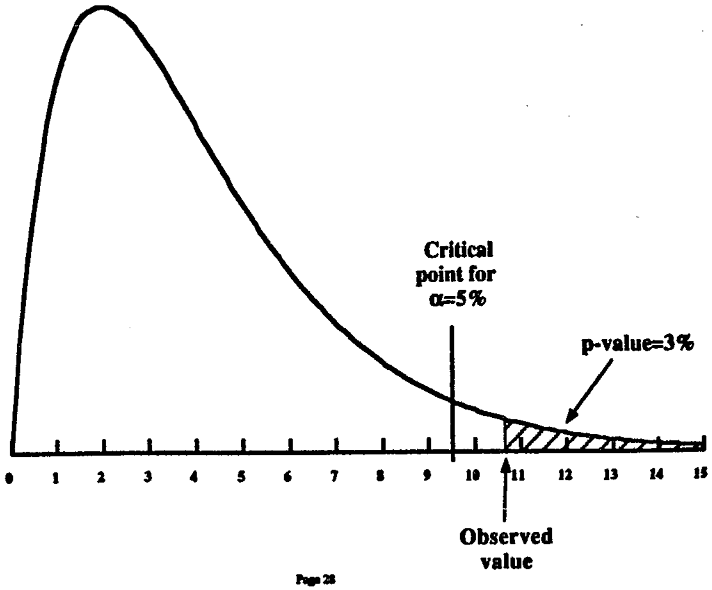
TYPE I vs. TYPE II ERROR



P-VALUES

- Alternative way to reports results of statistical tests
  - Before, significance level was fixed and critical point (CP) determined from α
  - H_0 was rejected if statistic was more extreme than CP
  - P-value is chance of seeing, under H_0, a statistic at least as extreme as the one found
- Rather than compute critical value, just compute p-value
  - If p-value is small enough, reject H_0
  - P-values are more precise than fixed significance levels

P-VALUES



Section 4

CHECKING ASSUMPTIONS

- Why? Statistical models approximate reality
- Before using a test, check underlying assumptions
  - Type of distribution (e.g., normal, log-normal)
  - Presence of seasonality or other dependence.
  - Homogeneity of variances across wells
  - Presence of outliers

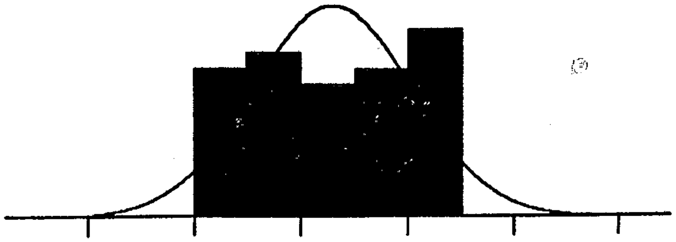
CHECKING DISTRIBUTIONS

- Choose an approximate model; eliminate bad models
- Choice of model very important
  - Normal versus lognormal data
  - Evidence of contamination may only be seen if logs of original data are tested
- Guidelines
  - If testing normality, run test on raw data
  - If testing lognormality, run test on logged data
  - Test background and compliance data separately (unless using residuals)

TESTING FOR NORMALITY

- Coefficient of Variation: CV = SD/Mean
  - Easy to calculate but not reliable, especially on logged data
  - For positive data, CV can indicate skewness
- Coefficient of Skewness, γ_1
  - Normal data ⇒ coefficient = 0
  - Non-normal data ⇒ coefficient positive or negative depending on distribution
  - $|\gamma_1| > 1$ ⇒ robustness of t-test deteriorates rapidly
- Chi-squared, χ^2, Test
  - Not most useful test of normality
  - May not detect departure from normality in tails
  - Recommend other tests

CHI-SQUARE TEST



TESTING FOR NORMALITY (cont.)

- Probability Plots (P-Plots)
  - Useful technique, but must supplement with numerical test
  - Straight line fit is good indication of normality
  - Help detect outliers, skewness
- Filliben's P-plot Correlation Coefficient
  - Excellent supplement to p-plot
  - Correlation between ordered data and normal quantiles
  - When correlation is too low, normal model is rejected
  - Note that p-plot correlations will always look high

TESTING FOR NORMALITY (cont.)

- Shapiro-Wilk Test
  - Can use with at least 3 but no more than 50 data points
  - Excellent numerical test of normality (similar to Filliben's test)
  - Powerful for detecting non-normality in tails
- Can be computed by hand easily
  - When W statistic is too low, normality is rejected
  - Interpretation of critical points
- Use Shapiro-Francia when n>50

CALCULATING SHAPIRO-WILK

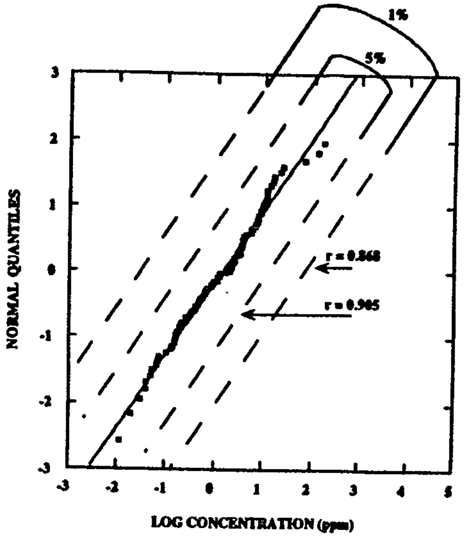
$$W = \left[\frac{b}{SD\sqrt{n-1}}\right]^2$$

Nickel Data:

$$W = \left[\frac{932.88}{(259.72)\sqrt{19}}\right]^2 = 0.679 < W_{.01,20} = 0.868$$

i	$x_{(i)}$	$x_{(n-i+1)}$	Difference	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	25.44	.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		

INTERPRETING SHAPIRO-WILK



OVERALL FRAMEWORK

- If data follow specific probability model (e.g., normal or lognormal)
  - Use parametric tests, because
  - Parametric tests more powerful for detecting differences than non-parametric tests
- If data are non-normal, non-lognormal
  - Find another transformation or better distribution, or
  - Use non-parametric tests (especially with many non-detects)

OVERALL FRAMEWORK (cont.)

- Normalizing transformations
  - Taking logs is one example
  - Others include square root, cube root, reciprocal
  - Done to get normality and to stabilize group variances
  - Transformed data can be tested for normality
- Other distributions: gamma, weibull, beta
- Nonparametric rank tests
  - Usually easier to compute, require fewer assumptions
  - Less powerful when data really follow a specific model
  - Often more powerful when data come from unknown model

DISTRIBUTION OF DATA		TYPE OF STATISTICAL PROCEDURE
Normal Lognormal	⇒	Parametric Test: • ANOVA • Tolerance Interval • Prediction Interval • Control Charts • t-test • Confidence Interval
Non-Normal Non-Lognormal	⇒	Non-parametric Test: • Kruskal-Wallis • Wilcoxon Rank-Sum • Non-parametric Intervals

INDEPENDENCE OF DATA

- Important for almost all statistical procedures
  - These tests depend on accurate estimate of variability
- Dependent samples exhibit less variability
  - Leads to underestimating true variance
  - Can severely affect results of statistical testing

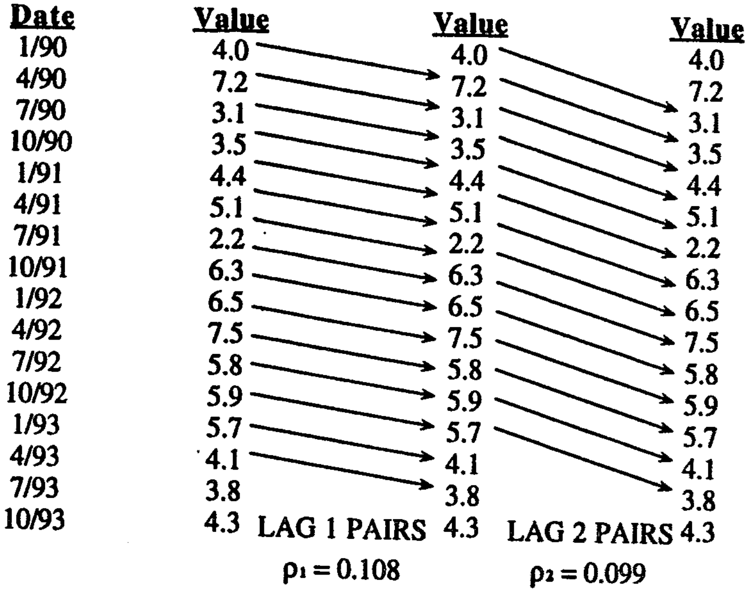
CURRENT GUIDANCE

- Try to ensure physical independence of samples
  - – Allow enough time between sampling episodes
  - – Only use ANOVA-type tests if groundwater velocity is above average
- If independence <u>cannot</u> be assured
  - – Use interval testing on single wells with fewer samples (e.g., 1 sample every 6 months)
- Important: physical independence does not guarantee <u>statistical independence</u>

TESTING FOR INDEPENDENCE

- Data not independent if adjacent sampling episodes are strongly correlated
- One method: estimate autocorrelations
  - – Standard correlation on single variable
  - – Data "pairs" are samples separated in time by a certain "lag"
  - – Examples: Lag 1 or Lag 2 autocorrelations
  - – If autocorrelations at all lags are zero, treat data as independent

AUTOCORRELATION

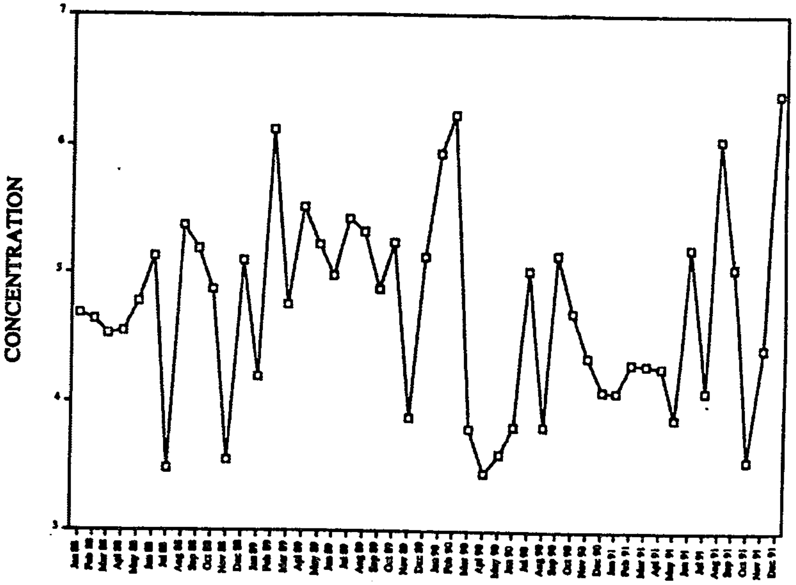


Date	Value	Value	Value
1/90	4.0	4.0	4.0
4/90	7.2	7.2	7.2
7/90	3.1	3.1	3.1
10/90	3.5	3.5	3.5
1/91	4.4	4.4	4.4
4/91	5.1	5.1	5.1
7/91	2.2	2.2	2.2
10/91	6.3	6.3	6.3
1/92	6.5	6.5	6.5
4/92	7.5	7.5	7.5
7/92	5.8	5.8	5.8
10/92	5.9	5.9	5.9
1/93	5.7	5.7	5.7
4/93	4.1	4.1	4.1
7/93	3.8	3.8	3.8
10/93	4.3	4.3	4.3

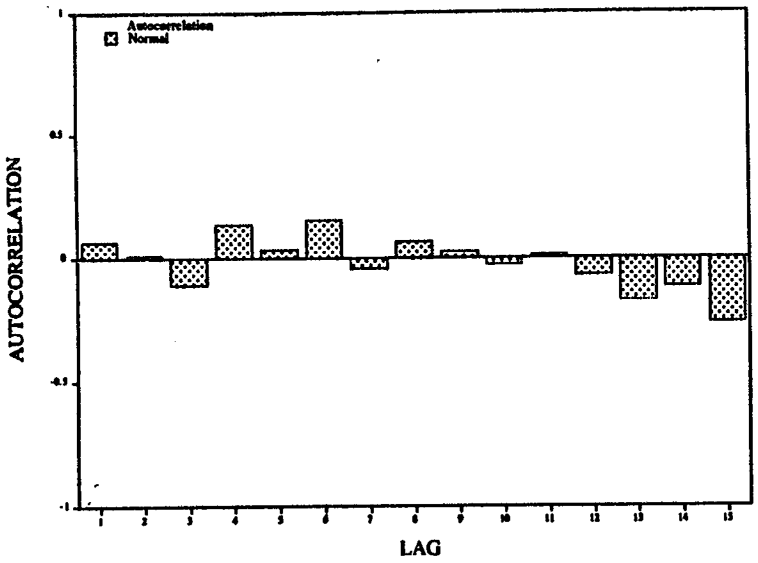
LAG 1 PAIRS $\rho_1 = 0.108$

LAG 2 PAIRS $\rho_2 = 0.099$

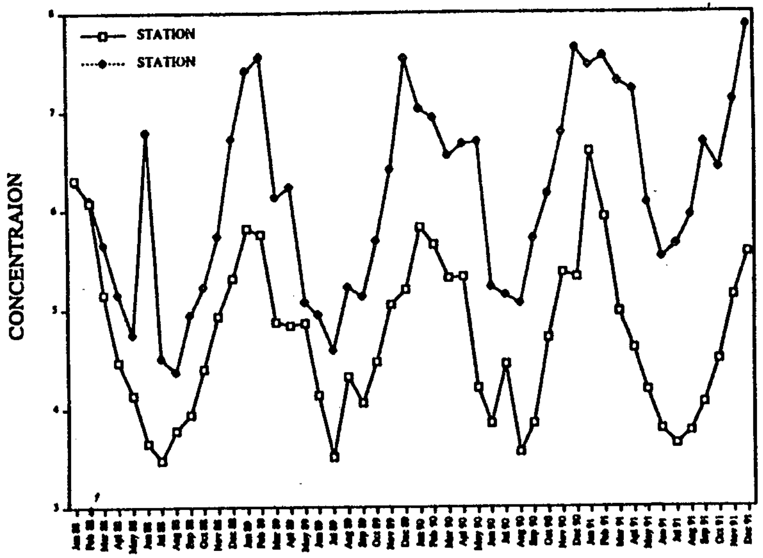
UNCORRELATED MONITORING DATA



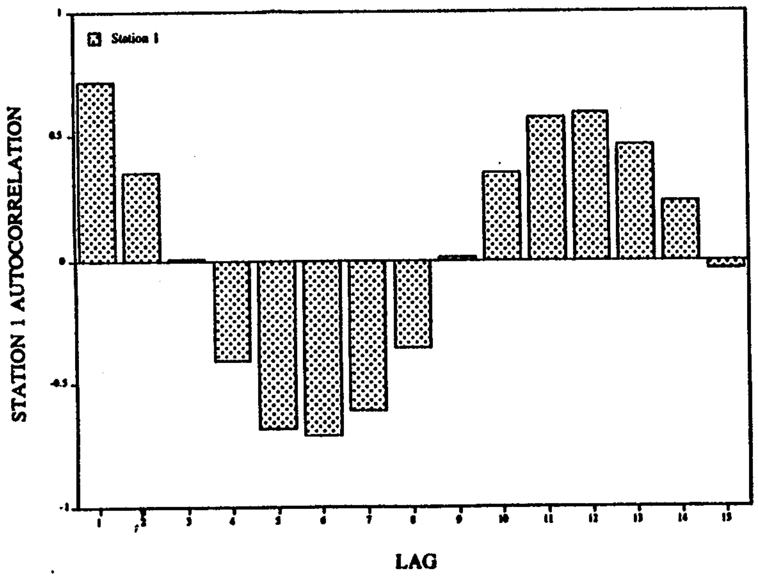
AUTOCORRELATIONS FOR RANDOM DATA



SEASONAL MONITORING DATA



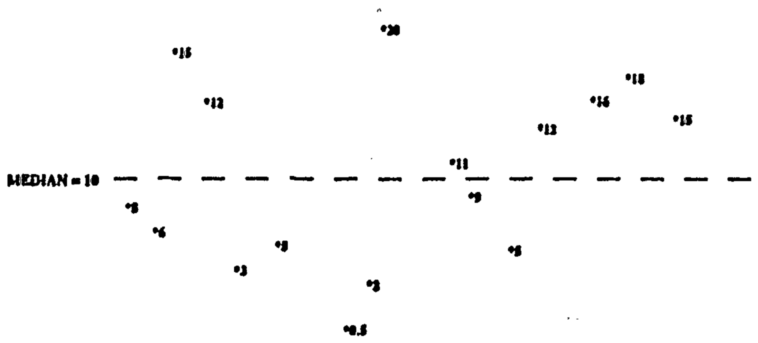
AUTOCORRELATIONS FOR SEASONAL DATA



TESTING INDEPENDENCE (cont.)

- Simpler test: runs count
  - Series of values should fluctuate randomly around median
  - Count number of runs above and below the median
  - If runs count is too low or too high, data not independent
- More powerful test: rank von Neumann ratio
  - Nonparametric test using ranks of data
  - First rank data, then compute von Neumann ratio
  - Reject independence if ratio is too low

RUNS COUNT



$(\underline{0\ 0}\ \underline{1\ 1}\ \underline{0\ 0\ 0\ 0}\ \underline{1\ 1}\ \underline{0\ 0}\ \underline{1\ 1\ 1\ 1})$

RUNS = 6

RANK VON NEUMANN RATIO

Date	Value	Rank
1/90	4.0	5
4/90	7.2	15
7/90	3.1	2
10/90	3.5	3
1/91	4.4	8
4/91	5.1	9
7/91	2.2	1
10/91	6.3	13
1/92	6.5	14
4/92	7.5	16
7/92	5.8	11
10/92	5.9	12
1/93	5.7	10
4/93	4.1	6
7/93	3.8	4
10/93	4.3	7

$$V = \frac{\sum_{i=2}^{n}(r_i - r_{i-1})^2}{n(n^2-1)/12}$$

$$= \frac{(10^2 + 13^2 + \cdots + 2^2 + 3^2)}{16(16^2-1)/12}$$

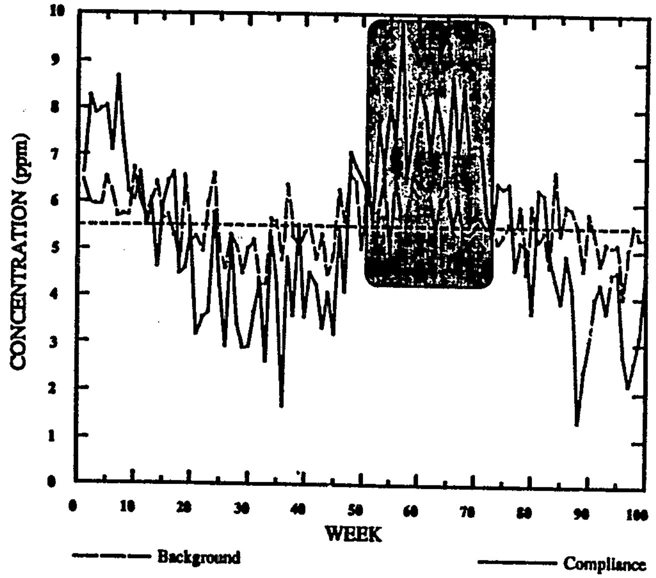
$$= 568/340 = 1.67$$

$V_{.05,16} = 1.21$ NS

SPECIAL TYPES OF DEPENDENCE

- Replicate Samples
  - Field or lab splits ⇒ measurements not independent!
  - Only measures sampling or analytical variability
- Serial or Temporal Correlation
  - Can be seen in time plots, control charts
  - Masks true variability if data from limited time period
  - Effect is minimal if patterns change on order of several years
  - Simple seasonal patterns can often be corrected
  - Have to adjust data before testing (e.g., deseasonalize data)

SEASONAL CORRELATION

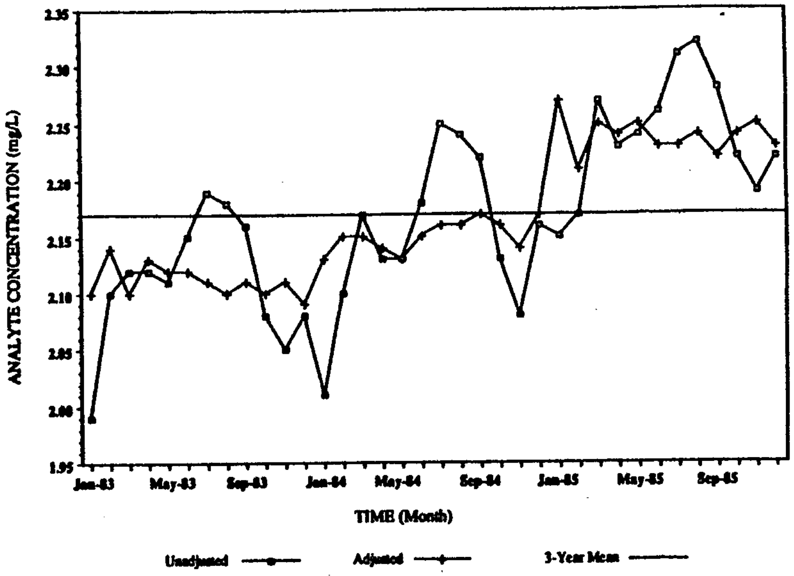


HOW TO DESEASONALIZE DATA

- Need large set of historical data
  - Determine length of full seasonal cycle
  - Divide into repeated time intervals at common points in the cycle
  - Example: 1 year cycle — monthly data over 3 years can be divided into 3 January points, 3 February points, etc.

TIME SERIES OF MONTHLY OBSERVATIONS

(Unadjusted, Adjusted, 3-year Mean)



HOW TO DESEASONALIZE DATA

(cont.)

- Calculate average of each repeated time interval

Month	'83	'84	'85	$\bar{X}_{mo}$
Jan	1.99	2.01	2.15	2.05
Feb	2.10	2.10	2.17	2.12
Mar	2.12	2.17	2.27	2.19
Apr	2.12	2.13	2.23	2.16
May	2.11	2.13	2.24	2.16
Jun	2.15	2.18	2.26	2.20
July	2.19	2.25	2.31	2.25
Aug	2.18	2.24	2.32	2.25
Sep	2.16	2.22	2.28	2.22
Oct	2.08	2.13	2.22	2.14
Nov	2.05	2.08	2.19	2.11
Dec	2.08	2.16	2.22	2.15
		3-year mean =		2.17

HOW TO DESEASONALIZE DATA

(cont.)

- Calculate seasonal adjustment for each data point

$$X'_{ij} = X_{ij} - \bar{X}_{i\bullet} + \bar{X}_{\bullet\bullet}$$

– Example:

$$\text{January}\begin{cases}1.99-2.05+2.17=2.11\\2.01-2.05+2.17=2.13\\2.15-2.05+2.17=2.27\end{cases}$$

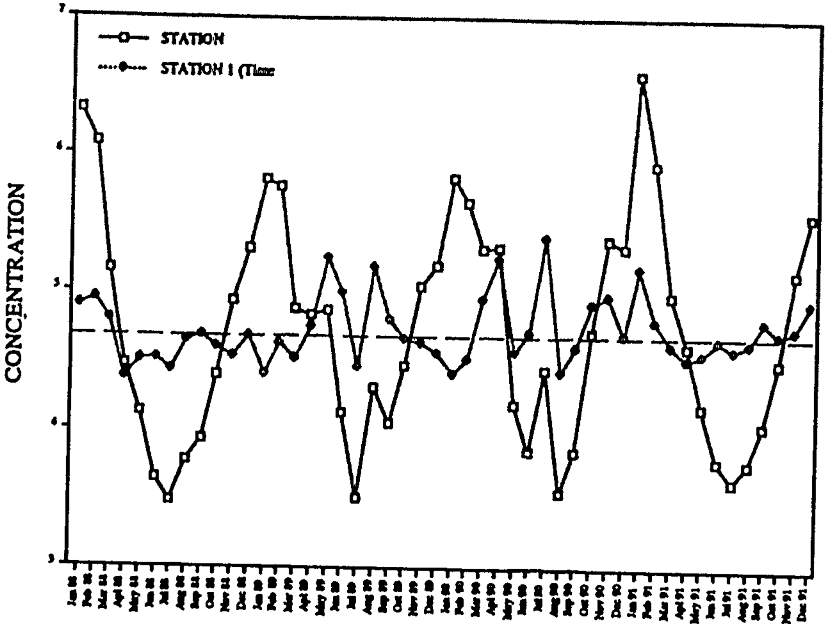
$$\text{June}\begin{cases}2.15-2.20+2.17=2.12\\2.18-2.20+2.17=2.15\\2.26-2.20+2.17=2.23\end{cases}$$

- Use adjusted values instead of original data in statistical procedures

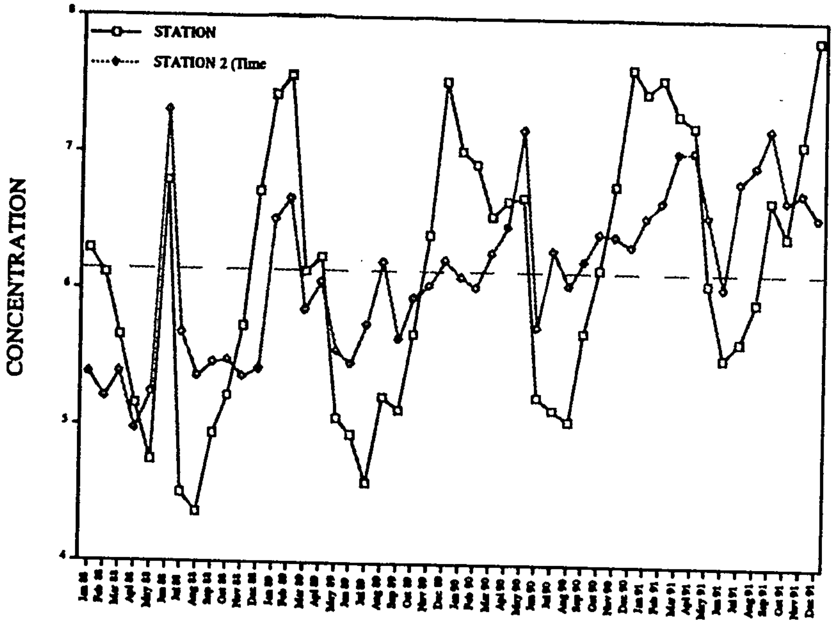
Seasonal Data

Station 1 Year	Number of Data Points n = 48 Month	Station 1	Station 2 Year	Number of Data Points n = 48 Month	Station 2
1	1	6.32	1	1	6.39
1	2	4.88	1	2	6.11
1	3	2.16	1	3	5.86
1	4	4.07	1	4	5.16
1	5	4.13	1	5	6.72
1	6	3.66	1	6	6.79
1	7	3.48	1	7	4.51
1	8	2.78	1	8	4.57
1	9	2.94	1	9	4.96
1	10	4.40	1	10	5.32
1	11	4.94	1	11	5.77
1	12	3.32	1	12	6.72
2	1	2.82	2	1	7.48
2	2	3.76	2	2	7.36
2	3	4.08	2	3	6.13
2	4	4.14	2	4	6.34
2	5	4.07	2	5	5.87
2	6	4.10	2	6	4.96
2	7	3.34	2	7	6.39
2	8	4.32	2	8	5.32
2	9	4.66	2	9	5.13
2	10	4.67	2	10	5.69
2	11	3.88	2	11	6.48
2	12	3.30	2	12	7.33
3	1	3.88	3	1	7.08
3	2	3.66	3	2	6.93
3	3	3.36	3	3	6.58
3	4	3.30	3	4	6.64
3	5	4.30	3	5	6.69
3	6	3.88	3	6	5.53
3	7	4.46	3	7	5.14
3	8	3.96	3	8	5.86
3	9	3.88	3	9	5.71
3	10	4.72	3	10	6.17
3	11	3.36	3	11	6.76
3	12	3.30	3	12	7.64
4	1	6.50	4	1	7.46
4	2	3.88	4	2	7.36
4	3	4.98	4	3	7.30
4	4	4.44	4	4	7.22
4	5	4.16	4	5	6.07
4	6	3.79	4	6	5.53
4	7	3.64	4	7	5.65
4	8	3.77	4	8	5.94
4	9	4.88	4	9	6.66
4	10	4.98	4	10	6.42
4	11	3.13	4	11	7.10
4	12	3.57	4	12	7.86

ADJUSTED STATION 1 MONITORING DATA



ADJUSTED STATION 2 MONITORING DATA



Section 5

Section 3

Section 4

TWO-SAMPLE COMPARISONS

- Use to compare background data from one or more wells against a single downgradient well
- May also be useful for early phases of monitoring with small number of observations at a few wells
  - Pool all downgradient wells into one group
  - Test will not pinpoint "guilty" well, but may be only option until additional data is collected
- Depending on data, two tests
  - Parametric: **Two-sample t-test**
  - Non-parametric: **Wilcoxon Rank-Sum test**

TWO-SAMPLE T-TEST

- Assumptions
  - Independent observations
  - Equal variances in both groups
  - Normal residuals
  - H_0: $\text{Mean}_{\text{Down}} = \text{Mean}_{\text{Bkg}}$
- To run:
  - Test residuals for normality and equal variance
  - Compute mean and SD in each group

TWO-SAMPLE T-TEST
(cont.)

- If variances are equal, compute t-statistic

$$t = (\text{Mean}_{\text{down}} - \text{Mean}_{\text{up}})/\text{SE}_{\text{diff}}$$

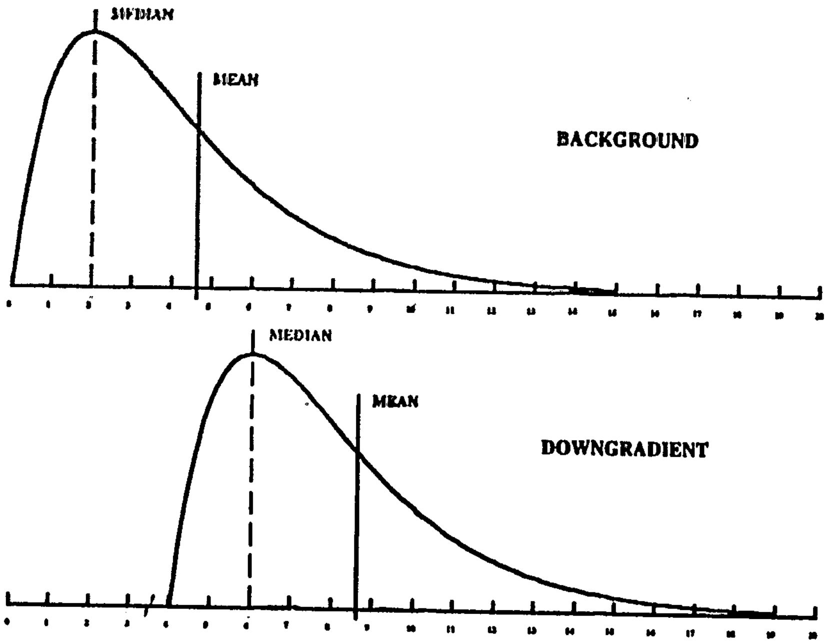
where SE_{diff} = standard error of mean difference

$$\text{SE}_{\text{diff}} = \sqrt{\left[\frac{(n_{\text{up}}-1)\text{SD}_{\text{up}}^2 + (n_{\text{down}}-1)\text{SD}_{\text{down}}^2}{(n_{\text{up}}+n_{\text{down}}-2)}\right]\left(\frac{1}{n_{\text{up}}}+\frac{1}{n_{\text{down}}}\right)}$$

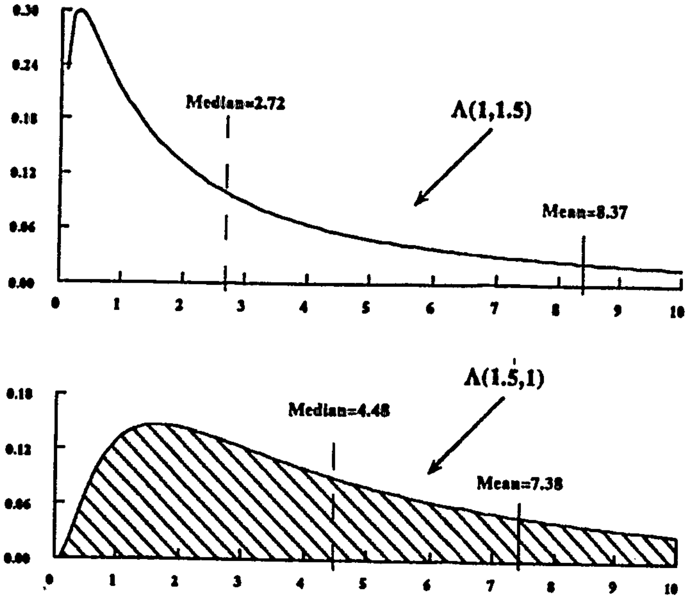
- Compare calculated t with one-sided critical point $t_c = t_{df,\alpha}$ where $df = (n_{\text{up}} + n_{\text{down}} - 2)$ and α = 1% or 5%

TWO-SAMPLE T-TEST
(cont.)

- If variances are not equal, use the CABF t-test
  - Built into GRITS/STAT
  - Key difference: degrees of freedom are adjusted
  - Eliminates need to test for equal variances
- If residuals are lognormal, run test on logged data
  - This case test for difference in medians not means
  - Difference in medians often, but not always, implies difference in means



MEAN OR MEDIAN?

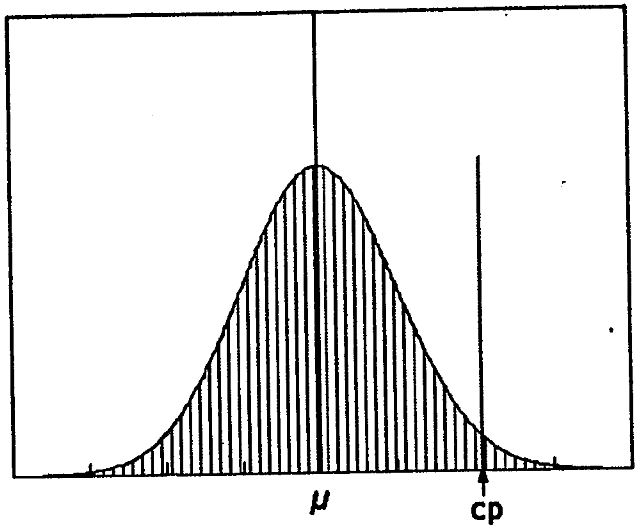


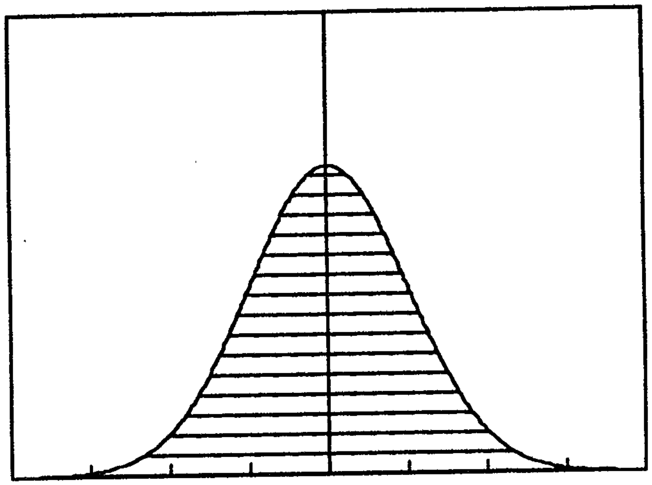
TWO-SAMPLE T-TEST
(cont.)

- Can calculate approximate power in three cases:
  - If $(\text{Mean}_{\text{down}} - \text{Mean}_{\text{up}}) = 0 \Rightarrow \text{power} = \alpha$
  - If $(\text{Mean}_{\text{down}} - \text{Mean}_{\text{up}}) = t_c \times SE_{\text{diff}} \Rightarrow \text{power} \approx 50\%$
  - If $(\text{Mean}_{\text{down}} - \text{Mean}_{\text{up}}) = 2(t_c \times SE_{\text{diff}}) \Rightarrow \text{power} \approx 100(1-\alpha)\%$
- Can find the chance that test will locate a difference in means at least as big as the right-hand side above
- Can use power results to determine necessary adjustments to sample size or α

TWO-SAMPLE T-TEST
(cont.)

- How does this work?
  - t-statistic: $t = (\text{Mean}_{\text{down}} - \text{Mean}_{\text{up}})/SE_{\text{diff}}$
  - $t \times SE_{\text{diff}} = \text{Mean}_{\text{down}} - \text{Mean}_{\text{up}}$
  - Critical point (CP) $= t_C \times SE_{\text{diff}} = \text{Mean}_{\text{down}} - \text{Mean}_{\text{up}}$
- Illustration





NOTES ON CABF T-TEST

- If there are more than 15% nondetects, use Wilcoxon rank-sum
- CABF still specified in some permits
- Valid procedure when assumptions satisfied
- Problems with CABF that led to revised regulations:
  - Assumptions substantially violated
  - Non-independent observations (lab or field replicates)
  - Not always most appropriate test (e.g., comparing >2 groups)

WILCOXON RANK-SUM TEST

- Use with many non-detects or when t-test assumptions not met
- Robust against non-normality or non-lognormality
- Basic algorithm
  - m background well samples, n compliance well samples
  - Rank all M = (n + m) data
  - W = Sum of compliance ranks minus $\frac{1}{2}n(n+1)$
  - Compare W with tabulated value
  - If W exceeds tabulated value ⇒ evidence of contamination

SPECIAL CONSIDERATIONS

- Sample size: at least 4 measurements per group
  - Otherwise statistical power will be too low
- Use normal approximation to W with continuity correction
  - By CLT, W is approximately normal
  - Built into GRITS/STAT
- If ties are present
  - Give each tie the average rank of the group of ties
  - Calculate W with an adjusted formula for the SD

WILCOXON RANK-SUM FORMULAS

$$W=\sum_{i=1}^{n}C_i-\frac{1}{2}n(n+1) \qquad E(W)=\frac{1}{2}mn$$

$$SD(W)=\sqrt{\frac{1}{12}mn(m+n+1)} \qquad Z=\frac{W-E(W)-.5}{SD(W)}$$

$Cp_{.01}=2.326$ $\qquad$ $Cp_{.05}=1.645$

WILCOXON RANK-SUM EXAMPLE

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

	Mean = 8.16
	SD = 2.55
			Mean = 15.00
			SD = 4.44

Ranks of Copper Concentration

	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
	$\Sigma R_j = 86$		$\Sigma R_j = 85$

COMPARISON WITH OTHER TESTS

- Comparison with Sign Test or Test of Proportions
  - Neither test accounts for magnitude of data
  - Wilcoxon more powerful than sign test
  - Wilcoxon usually more powerful than test of proportions
  - When proportion of non-detects is high (> 70%), test of proportions more powerful
  - But both lead to same conclusion, so just use Wilcoxon

Section 6

PARAMETRIC ANOVA

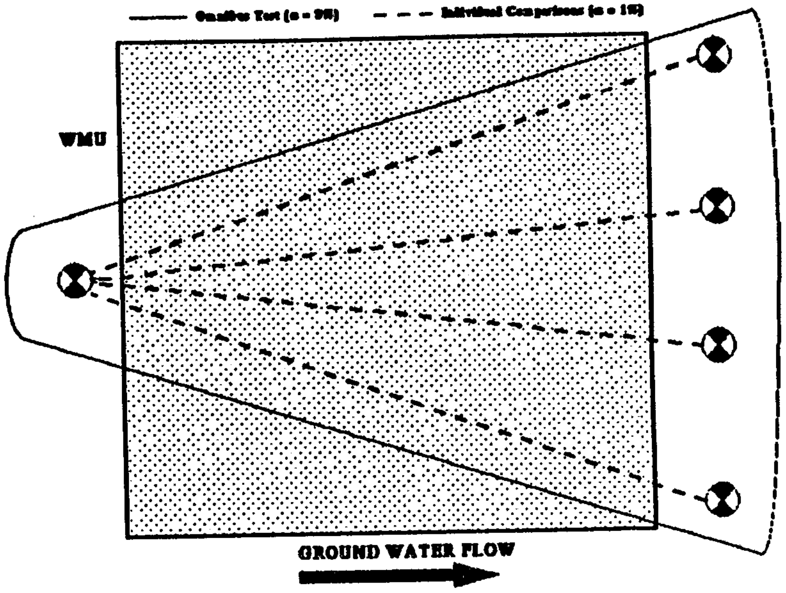
Use When:

- Testing several compliance wells simultaneously
- Comparing compliance data against background

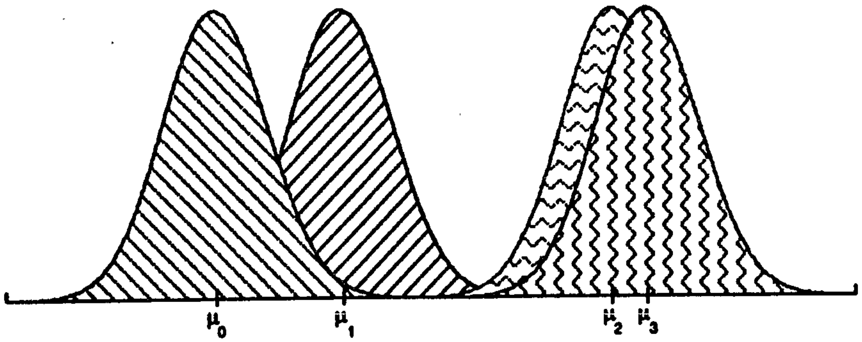
Objective: Test whether average concentration at any compliance well significantly exceeds mean background level

ANOVA is flexible and powerful when testing a small to moderate number of wells

MULTIPLE COMPARISONS



BEHIND ANOVA



BASIC ASSUMPTIONS

- Residuals are normally distributed
  - Each residual computed as measurement minus group mean
- Equal variances across well groups (i.e., homoscedasticity)
- Need to test these assumptions before running ANOVA
  - Calculate residuals and apply Shapiro-Wilk or Filliben's test
  - Use box plots and Levene's test to assess equal variances

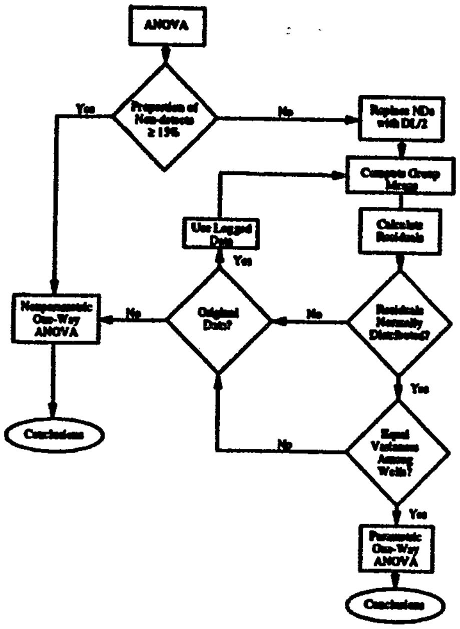
WHAT IF RESIDUALS ARE LOGNORMAL?

- Try re-running ANOVA on logged data
- Retest new residuals and data for normality and equal variances
- If assumptions check out, remember:
  - ANOVA on logged data tests for differences in medians in the original data
  - Difference in medians will typically imply a difference in means also, especially if equal variances holds on logged scale

DATA REQUIREMENTS

- Data must be classified into at least 3 groups
  - Background data makes up first group
  - Each compliance well is a separate group
- For ANOVA to work well, need:
  - <u>Minimum</u> of 3 to 4 samples per well
  - Total sample size, N, large enough so that $N-p \geq 5$, where p equals the number of groups

ANOVA FLOWCHART



CALCULATE RESIDUALS

Well #	Observations	Well Mean	Residuals
1	$X_{11}, X_{12}, X_{13}, X_{14}$	$\overline{X}_{1.}$	$R_{11}, R_{12}, R_{13}, R_{14}$
2	$X_{21}, X_{22}, X_{23}, X_{24}$	$\overline{X}_{2.}$	$R_{21}, R_{22}, R_{23}, R_{24}$
3	$X_{31}, X_{32}, X_{33}, X_{34}$	$\overline{X}_{3.}$	$R_{31}, R_{32}, R_{33}, R_{34}$
		$\overline{X}_{..}$	

$\overline{X}_{..}$ = grand (overall) mean

$R_{ij} = X_{ij} - \overline{X}_{i.}$ (measurement – group mean)

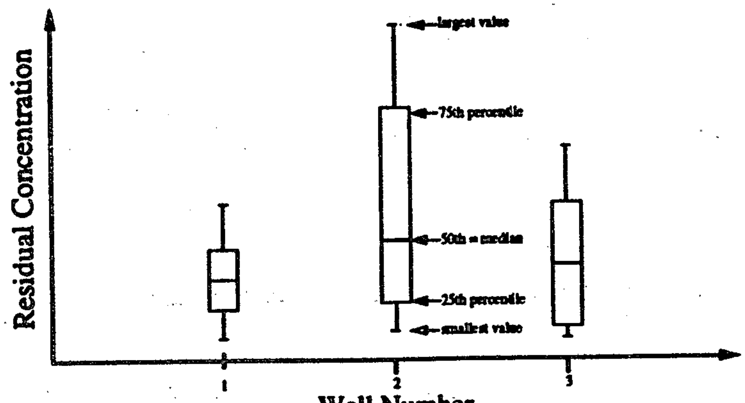
HOW TO CHECK RESIDUALS

- Testing normality
  - Pool all residuals together
  - Probability plot followed by Shapiro-Wilk or Filliben's test
- Senseless to test actual measurements
  - If group means are different, data may not *appear* normal
  - Residuals can be tested because means are removed
- Testing for equal variances
  - Box plot of each group
  - Compare box lengths
  - If too box lengths too different, run Levene's test

TESTING FOR EQUAL VARIANCES

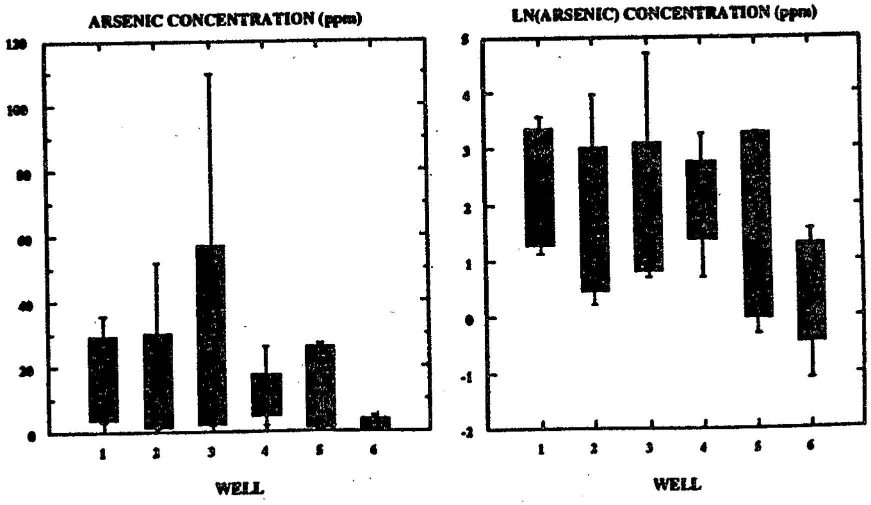
- Homogeneity of variances very important for ANOVA
  - More important than normality of residuals
  - Affects power of F-test to detect well mean differences
  - Ratio of largest to smallest group variances ≈ 4 ⇒ noticeable effect on power
  - Ratio exceeds 10 ⇒ severe effect on power
- Box-Plots
  - Quick way to visualize spread of data
  - Rule: if ratio of longest to shortest box length is more than 3, use Levene's test

BOX PLOTS



If ratio of longest to shortest box length < 3,
assume equal variances

BOX PLOTS OF WELL DATA



TESTING FOR EQUAL VARIANCES (cont.)

- More formal approach: Levene's Test
  - Not as sensitive to non-normality as Bartlett's test

- To run Levene's test:
  - Calculate residuals, R_{ij}, and take absolute values, $Z_{ij} = |R_{ij}|$
  - Perform ANOVA on these values, keeping groups as is
  - If F-test on the absolute residuals is significant, reject hypothesis of equal variances

WHAT NEXT?

- If both assumptions are valid, construct ANOVA table and compute the F-test

- If one or both assumptions does not pan out, particularly equal variances across groups, try an alternate transformation

- If no transformation works, try a nonparametric ANOVA (Kruskal-Wallis) on the ranks of the data
  - Particularly useful with a substantial fraction of non-detects

BASIC ANOVA TABLE

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F
Between Wells	SS_{wells}	p-1	$MS_{wells} = \frac{SS_{wells}}{(p-1)}$	$\frac{MS_{wells}}{MS_{error}}$
Error (within wells)	SS_{error}	N-p	$MS_{error} = \frac{SS_{error}}{(N-p)}$	
Total	SS_{total}	N-1		

Compare F with tabulated $F_{p-1,\, N-p,\, \alpha=5\%}$

ANOVA SUMS-OF-SQUARES

$$SS_{TOTAL} = (N-1)(SD)^2$$

$$SS_{ERROR} = \Sigma_{i=1}^{g} \Sigma_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i\right)^2 = \Sigma_{i=1}^{g} \Sigma_{j=1}^{n_i} R_{ij}^2$$

$$SS_{WELL} = \Sigma_{i=1}^{g} n_i \left(\bar{x}_i - \bar{x}\right)^2 = SS_{TOTAL} - SS_{ERROR}$$

DEGREES OF FREEDOM FOR F-TEST

Numerator = (#groups - 1)

$$F = \frac{MS_{WELLS}}{MS_{ERROR}}$$

Denominator = (N - #groups)

HOW TO INTERPRET ANOVA

- If F-test not significant, conclude there are no well differences
- F-test is significant, there is *probably* a significant difference
  - Always use common sense check: look at side-by-side box plots
  - To find out which, if any, compliance well exceeds background, do **post-hoc multiple comparisons**
    - Compute Bonferroni t-statistics (on each compliance well)
    - If less than 5 comparisons, let significance level equal α=5% divided by the number of compliance wells
    - If more than 5, perform each comparison at 1% significance level

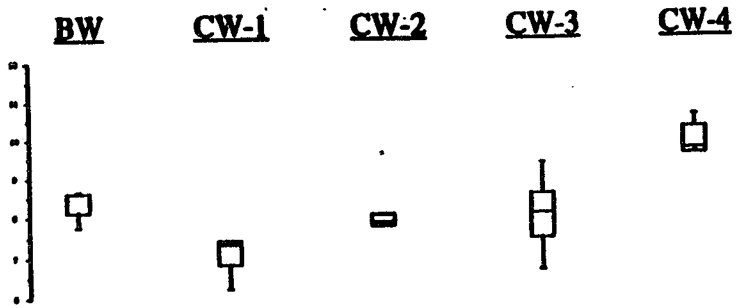
BONFERRONI t-STATISTICS

- Calculate average background level based on n_b samples
- Calculate (p-1) differences $\bar{X}_{i.} - \bar{X}_{Bg}$ $\quad i = 1,\dots,(p-1)$ compliance wells
- Divide these differences by SE : $SE_i = \left[MS_{error}\left(\frac{1}{n_b} + \frac{1}{n_i}\right)\right]^{1/2}$
- Use t table to find $t_{crit} = t_{N-p,\,1-\alpha^*}$ with $\alpha^*=.05/(p-1)$ or $\alpha^*=.01$ and N-p degrees of freedom, depending on p
- Compare each Bonferroni statistic t_i with t_{crit}
  - If t_i statistic is larger than t_{crit}, conclude that well i is out of compliance

UNUSUAL OUTCOMES

- Significant F-test does not *guarantee* the right kind of difference between two wells
  - Two compliance wells might differ from each other but not from background
  - Example
- Non-significant F-test can mask a significant pair-wise difference
  - Especially possible in larger network with a single contaminated well

UNUSUAL OUTCOME #1



BW	CW-1	CW-2	CW-3	CW-4
$\bar{X}=8.57$	$\bar{X}=7.32$	$\bar{X}=8.02$	$\bar{X}=7.96$	$\bar{X}=9.61$
s=1.045	s=1.003	s=0.960	s=1.031	s=0.966

$F = 3.67 > 2.87 = F_{.05,4,20} \Rightarrow$ Significant difference

$t_1 = -1.972 < 2.423 = t_{.0125,20} \Rightarrow$ NS
$t_2 = -0.868 \Rightarrow$ NS
$t_3 = -0.963 \Rightarrow$ NS
$t_4 = 1.641 \Rightarrow$ NS

Difference between
CW - 1 & CW - 2 is significant
$t = 3.614 > 2.423 = t_{.0125,20}$

UNUSUAL OUTCOME #2

Lead concentrations (ppm)

BW-1	CW-1	CW-2	CW-3	CW-4
7.5	8.0	7.8	6.6	9.65
7.1	9.6	7.7	8.0	10.25
7.5	8.1	9.6	7.4	8.05
9.75	7.5	7.0	9.3	10.55
8.0	6.9	8.0	8.5	9.55
$\overline{X}$=7.97	$\overline{X}$=8.02	$\overline{X}$=8.02	$\overline{X}$=7.96	$\overline{X}$=9.61
S=1.045	S=1.003	S=0.960	S=1.031	S=0.966

EXAMPLE TABLE

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F
Between Wells	10.48	4	2.62	2.61
Error (within wells)	20.07	20	1.00	
Total	30.55	24		

Since $F < F_{4,20,\alpha=5\%} = 2.87$, no significant difference is found

ALTERNATE ANOVA: DUNNETT'S

- Method to compare single group against k other data sets
  - No overall F-test, but t-tests of each compliance well against background
  - Dunnett's test uses special critical points
- When assumptions met, will be more sensitive to concentration increases over background
  - Why? Unlike F-test, Dunnett's method looks at smaller set of comparisons

ASSUMPTIONS OF DUNNETT'S TEST

- Residuals are normally distributed
- Equal variances across well groups
- Equal sample sizes in all groups, including background
  - Special critical points only exist for this case
  - Interpolation scheme exists to get critical points when background sample size is larger than any compliance well, but only for two-sided comparisons

BASIC ALGORITHM

- For each compliance well, compute t-statistic

$$t_i = \frac{\sqrt{n}(\bar{y}_i - \bar{y}_0)}{s\sqrt{2}}$$

  - s^2 represents the common pooled variance

$$s^2 = \sum_{i=0}^{k} s_i^2 \Big/ (k+1)$$

- Compare each t-statistic against special critical point at significance level α=.05, with d.f. equal to k and ν=(k+1)(n−1)
  - Each t-statistic greater than the critical value represents a significant increase over background

DUNNETT'S EXAMPLE

- Applying Dunnett's test to unusual example #2
  - Compute the pooled variance, s^2=1.004
  - Compute Dunnett's t-statistics for one-sided comparisons of each compliance well to BW-1
  - t_1= 0.0789
t_2= 0.0789
t_3= −0.0158
t_4= 2.5879
- Look up critical value in Dunnett's table with k=4 and ν=20 degrees of freedom: $d_{.05,4,20}$=2.30
  - Since $t_4 > d_{.05,4,20}$, significant difference is found at CW-4

NON-PARAMETRIC ANOVA (KRUSKAL-WALLIS)

- Use when parametric ANOVA does not apply
  - ANOVA assumptions grossly violated
  - Recommended whenever fraction of NDs > 15%
- Does not make assumptions about underlying distribution
  - Still assumes approximately equal variances
- When comparing 2 groups ⇒ use Wilcoxon rank-sum test
- When comparing several groups ⇒ use Kruskal-Wallis test

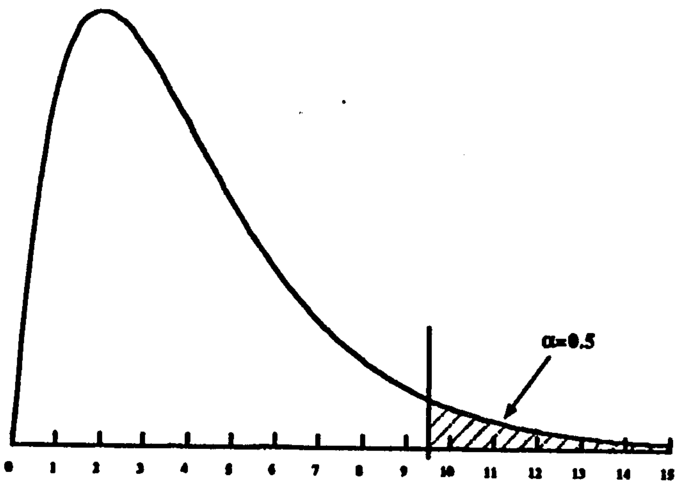
KRUSKAL-WALLIS ALGORITHM

- Compute ranks of combined data set
- Compute sum of ranks and average rank within each group
  - K groups: 1 background, (K-1) compliance wells
- Calculate **Kruskal-Wallis** test statistic, H

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

- Why is H computed?
  - Because it has a known distribution, the chi-square

CHI-SQUARE DENSITY



INTERPRETING KRUSKAL-WALLIS

- Compare H with appropriate chi-squared critical value, χ^2, with degrees of freedom = K-1
- If H is less than χ^2, conclude there are no pairwise differences in medians
- If H is greater than $\chi^2 \Rightarrow$ significant difference between at least two well groups
  - To locate the difference(s), do post-hoc multiple comparisons

NON-PARAMETRIC MULTIPLE COMPARISONS

- Individual comparisons done between background (group 1) and each compliance well
- Compute difference between average rank of each group and average rank of background
- Divide by approximate standard error:

$$SE_i = \left[\frac{N(N+1)}{12}\right]^{1/2}\left[\frac{1}{n_l}+\frac{1}{n_i}\right]^{1/2}$$

  for i = 2,..., K to get ratio z_i
- Compare each ratio z_i to $Z_{\alpha/(K-1)}$, the upper $\alpha/(K-1)$-percentile from N(0,1) using α=5%

KRUSKAL-WALLIS SPECIAL CONSIDERATIONS

- Need at least 3–4 measurements per well
- If more than 5 compliance wells, perform individual comparisons using $Z_{(\alpha=0.01)}$ instead of $Z_{\alpha/(K-1)}$
- If ties are present (some values are numerically equal)
  - Give each set of ties the average rank of that set
  - Calculate H´ (H statistic adjusted for ties)
- To test equal variances, run side-by-side boxplots of ranked values

EXAMPLE DATA FOR ONE-WAY NON-PARAMETRIC ANOVA

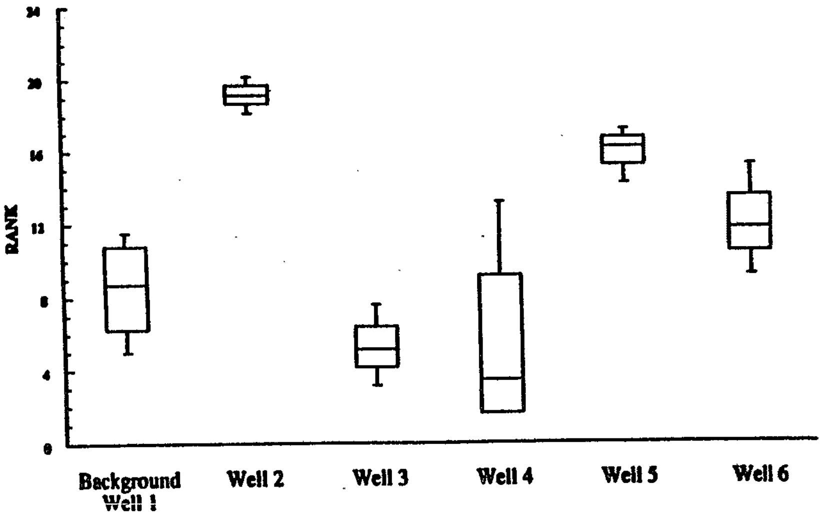
Date	Background Well 1	Compliance Wells Benzene concentration, ppm (Rank)				
		Well 2	Well 3	Well 4	Well 5	Well 6
Jan 1	1.7 (10)	11.0 (20)	1.3 (5)	<1 (1.5)	4.9 (17)	1.6 (9)
Feb 1	1.9 (11.5)	8.0 (18)	1.2 (3)	1.3 (5)	3.7 (16)	2.5 (15)
Mar 1	1.5 (7.5)	9.5 (19)	1.5 (7.5)	<1 (1.5)	2.3 (14)	1.9 (11.5)
Apr 1	1.3 (5)			2.2 (13)		
	$n_1 = 4$	$n_2 = 3$	$n_3 = 3$	$n_4 = 4$	$n_5 = 3$	$n_6 = 3$
Sum of ranks	$R_1 = 34$	$R_2 = 57$	$R_3 = 15.5$	$R_4 = 21$	$R_5 = 47$	$R_6 = 35.5$
Average rank	$\bar{R}_1 = 8.5$	$\bar{R}_2 = 19$	$\bar{R}_3 = 5.17$	$\bar{R}_4 = 5.25$	$\bar{R}_5 = 15.67$	$\bar{R}_6 = 11.83$

K = 6, total number of wells
N = 20, total number of observation

RANKING TIES

Order	Concentration	Rank	
1	<1	1.5	$\Rightarrow \frac{1}{2}(1+2)$
2	<1	1.5	
3	1.2	3	
4	1.3	5	$\Rightarrow \frac{1}{3}(4+5+6)$
5	1.3	5	
6	1.3	5	
7	1.5	7.5	$\Rightarrow \frac{1}{2}(7+8)$
8	1.5	7.5	
9	1.6	9	

CHECKING RANK VARIANCES



COMPUTING KRUSKAL-WALLIS

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

$$H = \left[\frac{12}{20(21)}\left(\frac{34^2}{4}+\frac{57^2}{3}+\frac{15.5^2}{3}+\frac{21^2}{4}+\frac{47^2}{3}+\frac{35.5^2}{3}\right)\right] - 3(21) = 77.68 - 63 = 14.68$$

⇒ Adjust for ties (4 groups)

$$H' = \frac{H}{1 - \Sigma_{i=1}^{g} \frac{t_i^3 - t_i}{N^3 - N}} \qquad N = 20$$

$$H' = \frac{14.68}{1 - \left\{3\left[\frac{2^3-2}{20^3-20}\right] + \left[\frac{3^3-3}{20^3-20}\right]\right\}} = 14.76$$

COMPUTING KRUSKAL-WALLIS (cont.)

- Compute degrees of freedom = # wells–1 = 6 – 1 = 5
- Look up critical value, $\chi^2_{5,.05}$ = 11.07 from Chi-Square table
- Since H′ = 14.76 > 11.07, have significant difference
  - – Need to test each compliance well versus background in pairwise comparisons

KRUSKAL-WALLIS PAIRWISE COMPARISONS

- Compute average rank differences and z_i ratios

CW	Ave. Rank Diff.	SE	z_i	z_{crit}
2	19 – 8.5 = 10.5	4.52	2.324	2.324
3	5.2 – 8.5 = –3.3	4.52	–.730	2.324
4	5.2 – 8.5 = –3.3	4.18	–.789	2.324
5	15.7 – 8.5 = 7.2	4.52	1.593	2.324
6	11.8 – 8.5 = 3.3	4.52	.730	2.324

- One borderline significant difference at CW #2
- Other CW's are not significantly different from background
  - – Differences do exist between other compliance wells

Section 7

CONTROL CHARTS

- Alternative method for:
  - Intra-well comparisons during detection monitoring, or
  - Comparison to historically-monitored background wells
- Graphical tool to track data and discern
  - Rapid changes
  - Long term trends
- Data plotted on a control chart in cumulative fashion
  - New data added to chart as it comes in
  - Possible to see overall historical picture

CONTROL CHART LIMITATIONS

- Intrawell comparisons eliminate worry about spatial variation
- However:
  - Can only use on parameters with high detection rates (i.e., few NDs)
  - Can only construct on initially uncontaminated wells
  - Why? Need "clean" baseline measurement; otherwise, contamination hard to identify
- Modified control charts can be designed to monitor decreases in already contaminated wells

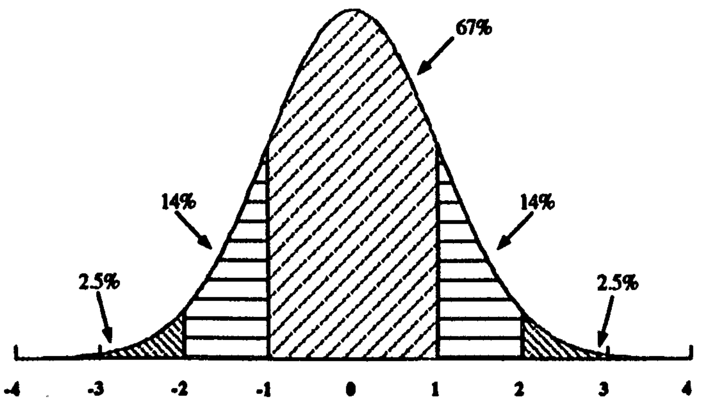
HOW DO CONTROL CHARTS WORK?

- Need initial data to estimate **mean** (μ) and SD (σ)
  - Initial data are not plotted
  - At least 8 independent samples
- If baseline data are not independent, first deseasonalize
- All future samples are standardized with respect to μ and σ
  - Mean of each sampling period is transformed

$$\overline{X}_i \rightarrow Z_i = \sqrt{n_i}\left(\overline{X}_i - \mu\right) / \sigma$$

- Plot Z_i's over time

STANDARD NORMAL DISTRIBUTION N(0,1)



HOW DO CONTROL CHARTS WORK (cont.)

To detect slow increase:

- Calculate cumulative sums, S_i

  $S_i = \max\{0, (Z_i - k) + S_{i-1}\}$

  where

  $S_0 = 0$

  $S_1 = \max\{0,(Z_1 - k) + 0\}$ at time T_1

  .

  .

  .

  .

  $S_t = \max\{0, (Z_t - k) + S_{t-1}\}$ at time T_t

- Plot S_i's over time on chart

CONTROL CHART THRESHOLDS

- Need statistical criterion to help detect contamination

- 3 parameters are necessary

h	=	decision internal value
k	=	reference control limit
SCL	=	Shewhart Control Limit

- Recommended Values (Starks, 1988)

k	=	1
h	=	5
SCL	=	4.5

CONTROL CHART THRESHOLDS (cont.)

- h is used for testing CUSUM

- SCL is used for testing individual standardized means

- Process "out of control" in one of two ways:

  1. Standardized mean, Z_i, exceeds the SCL

  2. The CUSUM of the standardized mean exceeds threshold, h

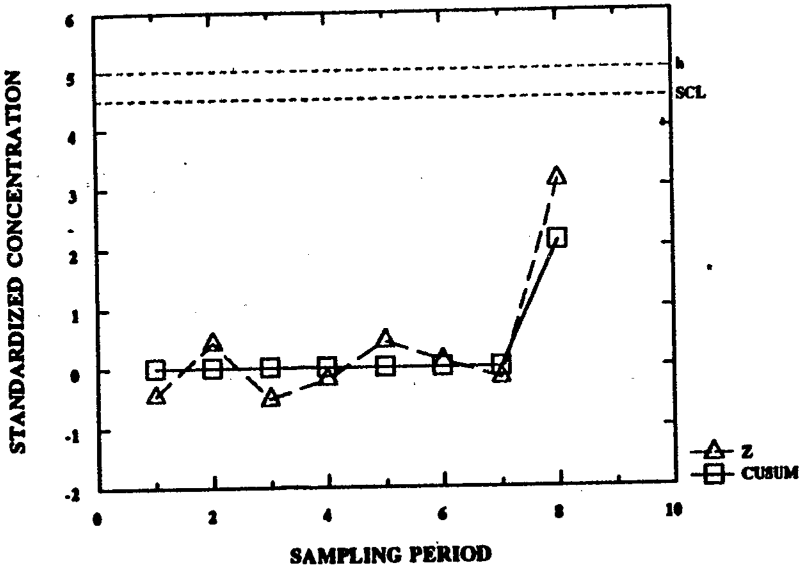
INTERPRETING CONTROL CHARTS

- Control chart declared "out-of-control" when sample data become too large relative to baseline parameters
  - Idea: as contamination occurs, true baseline values will rise
  - Standardizing by *original* baseline, however, will cause the plotted values to climb over fixed thresholds

- Thresholds are set so that a crossing identifies a significant increase

CONTROL CHART FOR NICKEL

MU = 27ppb SIGMA = 25ppb



UNDERLYING ASSUMPTIONS

- Data are normally distributed

  – Test baseline data for normality; used logged data if necessary

- Data are independently distributed

  – Remove seasonality if present

- Baseline parameters (mean, SD) reflect an "in control" process

– Update when more data become available and no contamination present

– Can run t-test to compare newer data with original baseline data before updating

SUMMARY

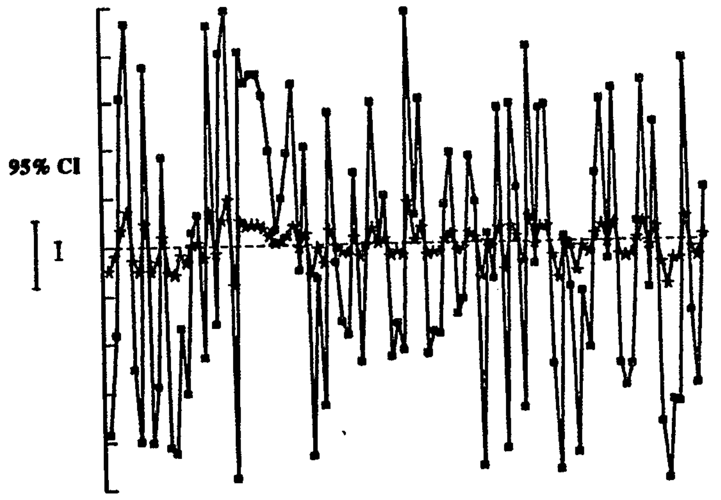
SITUATION	APPROACH
1. Few Obs. or Wells; Initial monitoring	Compare Background to Compliance 2-Sample method: t-test or Wilcoxon Rank-Sum
2. Many Wells; Several Obs. per well; No time effects	ANOVA Compare several wells: Parametric or Kruskal-Wallis test
3. Extensive Data Over Time; Seasonality or Spatial Variability	Intra-well Comparison Track a well's data over time: Control Charts or Prediction Limits

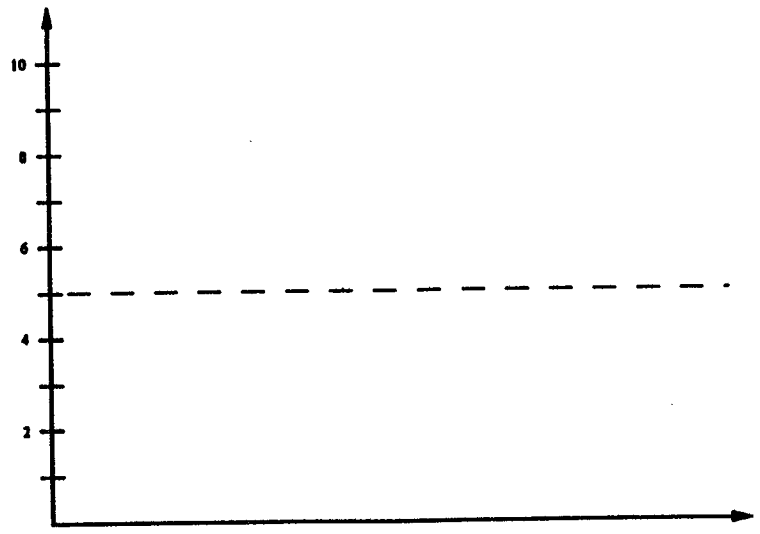
Section 8

OVERVIEW: STATISTICAL INTERVALS

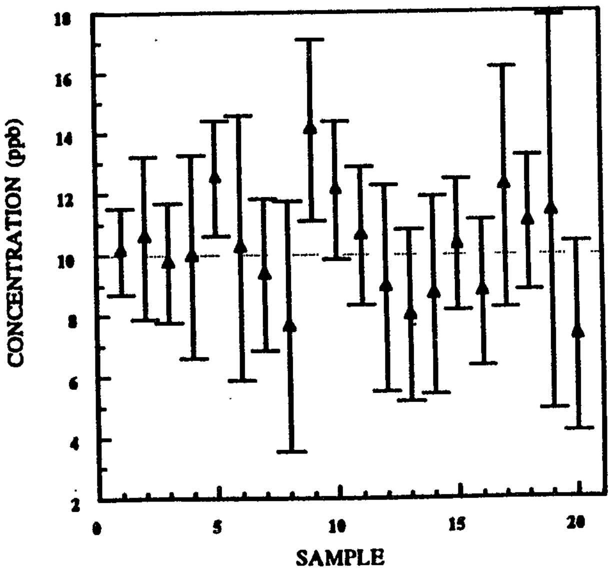
- Goals:
  - Estimate bounds on population characteristic
  - Predict range of future sample values
- Want an interval because point estimates say nothing about variability
  - Example
  - In-class experiment
- Random intervals constructed in same way will vary from dataset to dataset, even though underlying characteristic does not change

VARIABILITY IS CRUCIAL





REPEATED CONFIDENCE INTERVALS WHEN $\mu = 10$ (ppb)



HOW ARE INTERVALS USED?

- Not usually interested in mere estimation; instead want to
  - compare data to background
  - compare data to a groundwater protection standard (GWPS)
- Can use intervals to do "back-handed" hypothesis testing
  - Do downgradient data exceed GWPS?
  - Are downgradient measurements too different from estimated background characteristic?

CONFIDENCE AND COVERAGE

- Confidence level: how often an interval will contain/cover the population characteristic in repeated experiments
  - Error occurs when population target (e.g., mean) not covered by interval (happens $\alpha\%$ of the time)
  - Error of "missed containment"
  - Confidence level = $1-\alpha$
- Example of confidence interval for background mean

$$X \pm t_{n-1,\,\alpha} \frac{SD}{\sqrt{n}}$$

WIDTH OF INTERVAL

- Indicates amount of potential error associated with point estimate
- Width depends on 3 factors:
  - Standard deviation
  - Confidence level $(1-\alpha)$
  - Sample size
- To reduce width, either increase sample size or lower the confidence level

ASSUMPTIONS FOR PARAMETRIC INTERVALS

- Usual formulas based on having normal data
- If data are lognormal, method of attack depends on interval
  - Confidence intervals for the mean require special formulas
  - Other intervals use standard formulas, but must be computed on logged data
  - After constructing interval on logged scaled, retransform interval limits to original scale
  - All cases require computation of mean and SD for logged data
- If normality assumption violated, can try non-parametric interval

COMPUTING STATISTICAL INTERVALS

- General formula (except for lognormal CIs on the mean)

$$\bar{X} \pm K \cdot SD$$

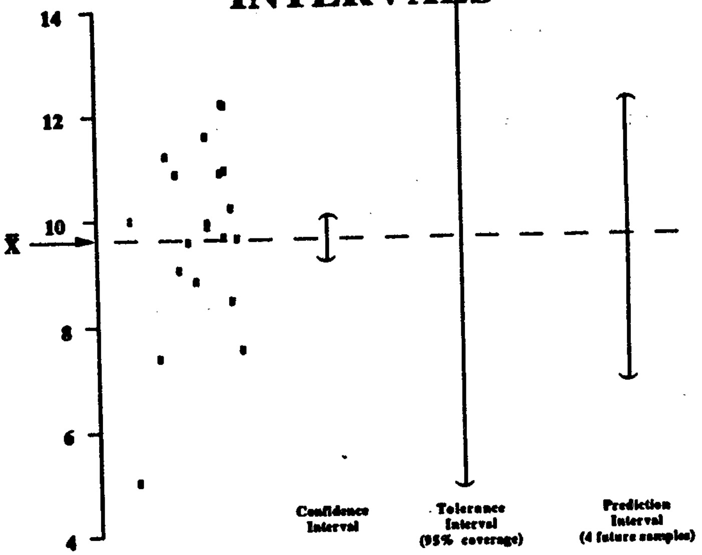
- Necessary components

  $\bar{X}$ = sample mean

  SD = sample standard deviation

  K = depends on interval type, sample size, and confidence level

STATISTICAL INTERVALS



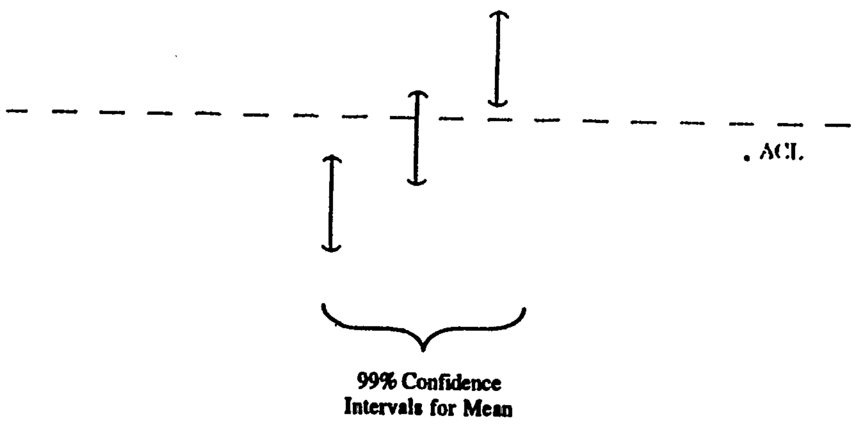
CONFIDENCE INTERVALS

- Two types used in ground-water monitoring
  - Interval containing the mean concentration
  - Interval containing an upper percentile (usually the 95th)
- Each type useful for particular scenario:
  - Compare CI for mean to ACL determined from average background data
  - Compare CI for upper percentile to risk-based MCLs or ACLs

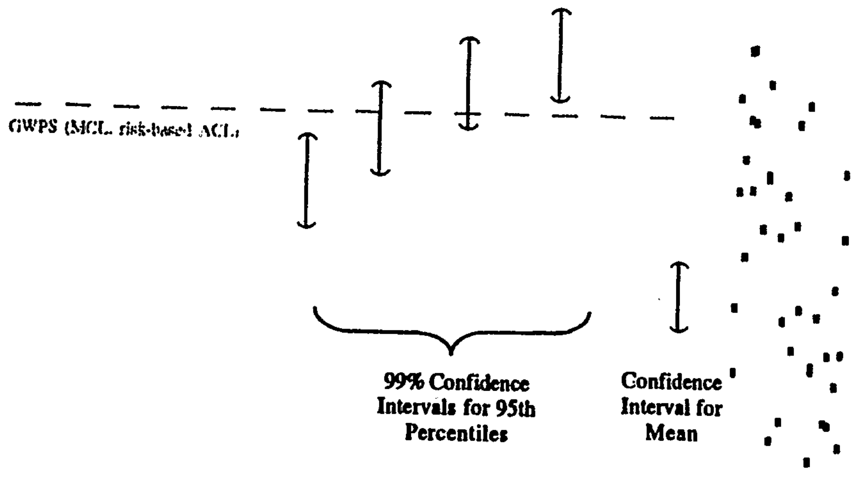
CONFIDENCE INTERVAL TESTS

- Only need to compute lower 99% confidence bound
  - If lower CI bound above GWPS, test triggers
- If comparing to ACL based on mean background data:
  - Compute CI containing mean of compliance point data
  - If ACL is less than lower CI limit, have possible violation
- If comparing to risk-based MCL or ACL
  - Compute CI containing upper (95th) percentile of compliance point data
  - If GWPS is lower than CI limit, have possible violation

COMPLIANCE MONITORING:
STANDARDS BASED ON BACKGROUND



COMPLIANCE MONITORING:
RISK-BASED STANDARDS



COMPUTING CONFIDENCE INTERVALS

- Simplest case: CI containing mean of normal data
  - Compute $\overline{X}$ and SD
  - Let $\kappa = t_{n-1,\ .01} \times \frac{1}{\sqrt{n}}$ for lower 99% confidence bound
  - Lower bound = $\overline{X} - \kappa \cdot SD$
- CI for mean of lognormal data
  - Compute mean ($\bar{y}$) and SD (s_y) of logged data
  - Use Land's (1971) formula for one-sided lower CI limit

$$LL_{\alpha} = \exp\left(\bar{y} + 0.5s_y^2 + \frac{s_y H_{\alpha}}{\sqrt{n-1}}\right)$$

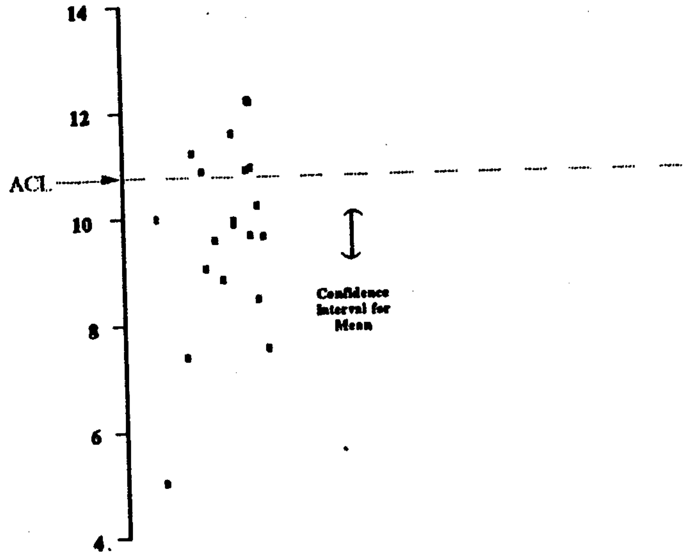
COMPUTING CONFIDENCE INTERVALS

- CI containing upper 95th percentile of normal data
  - Compute $\overline{X}$ and SD
  - Find κ from table in Hahn and Meeker (1993)
  - Lower bound = $\overline{X} + \kappa \cdot SD$
- CI containing upper 95th percentile of lognormal data
  - Compute mean ($\bar{y}$) and SD (s_y) of logged data
  - Again find κ from Hahn and Meeker table as before and compute lower 99% confidence limit
  - Exponentiate lower limit to find CI bound for the original data

NOTES ON CONFIDENCE INTERVALS

- Good estimate of variability is crucial
  - Must have at least 4 samples; recommend 8-10
- Caution regarding usual CI for the mean
  - Only designed to put bounds on the average level
  - Does not estimate range of individual concentrations or any upper percentile
- If CI for the mean is compared to risk-based MCL or ACL
  - Could have significant fraction of observations above GWPS, yet
  - Well deemed "clean" on basis of confidence interval test

CONFIDENCE INTERVALS



TOLERANCE INTERVALS

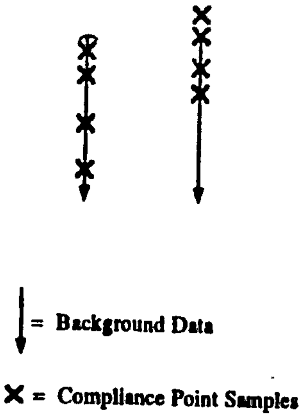
- Places limits on the likely range of possible individual measurements (i.e., the underlying population)
- Two parameters
  - Confidence level = 1-α
  - Coverage Coefficient (γ) = fraction of population bounded by the tolerance limits
- Only need upper 1-sided tolerance bound
  - Usually 95% confidence, 95% coverage

USING TOLERANCE INTERVALS

- Detection Monitoring
  - Compute upper tolerance limit (TL) on background data
  - Compare individual downgradient observations to upper TL
  - If any sample falls above TL (with γ=95% coverage), test triggers
- Note on larger compliance point sample sizes (>20)
  - Data from several compliance wells are often compared against a single upper tolerance limit
  - Expect 1 of every 20 samples to fail upper TL even with no contamination
  - In this case, increase γ or make provision for retesting

DETECTION MONITORING:

TOLERANCE LIMITS



COMPUTING TOLERANCE LIMITS

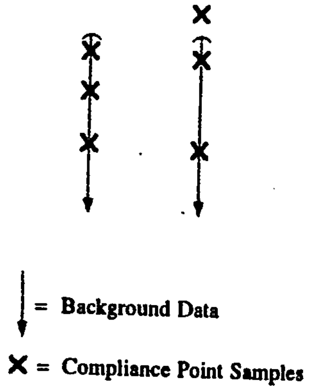
- For normal data:
  - Calculate $\bar{X}$ and SD on background data
  - Look up factor K in Hahn/Meeker table for one-sided tolerance limit with 95% coverage and 95% confidence
  - Compute upper TL as $\bar{X}+\kappa \cdot SD$
- For lognormal data:
  - Calculate mean($\bar{y}$) and SD (s_y) on logged background data
  - Compute upper TL as before, then exponentiate upper limit to make comparisons with original data
- Minimum data: again recommend 8-10 background samples

PREDICTION INTERVALS

- Estimates upper bound on next k future samples
  - Upper TL puts bound on fraction γ of *all* future samples
  - k can be as small as one; useful with limited compliance data
- In detection monitoring
  - Compute upper prediction limit (PL) on background data
  - If any of k compliance samples exceeds PL, test triggers
- Intrawell comparisons
  - Compute upper PL on past data
  - If any future sample exceeds upper PL, test triggers

DETECTION MONITORING:

PREDICTION LIMITS



COMPUTING PREDICTION INTERVALS

- For normal data:
  - Compute sample mean and SD
  - Determine number of samples (k) to be collected in next period
  - Calculate factor κ using the formula

$$\kappa = t_{.05/k,n-1} \times \sqrt{1 + \frac{1}{n}}$$

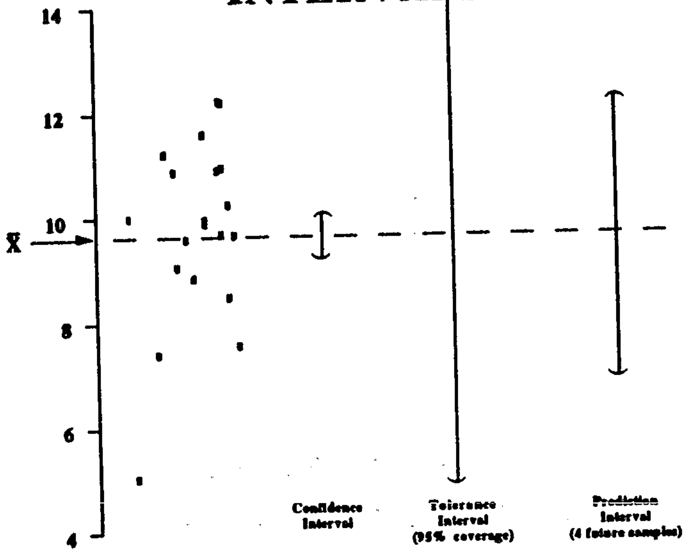
where n=# background data

  - Compute upper PL as $\overline{X} + \kappa \cdot SD$
- For lognormal data, compute upper PL on logged data
  - Exponentiate upper PL to make comparisons on original scale

NOTES ON PREDICTION LIMITS

- Prediction limits usually wider than confidence intervals, but shorter than tolerance limits
  - Why?
- Advantage: can be designed to predict one compliance sample per testing period
  - Not always feasible to collect > 1 independent sample
- Especially useful if the number of samples to be collected at each compliance point is identical and known in advance
  - Only need to compute 1 prediction limit from background

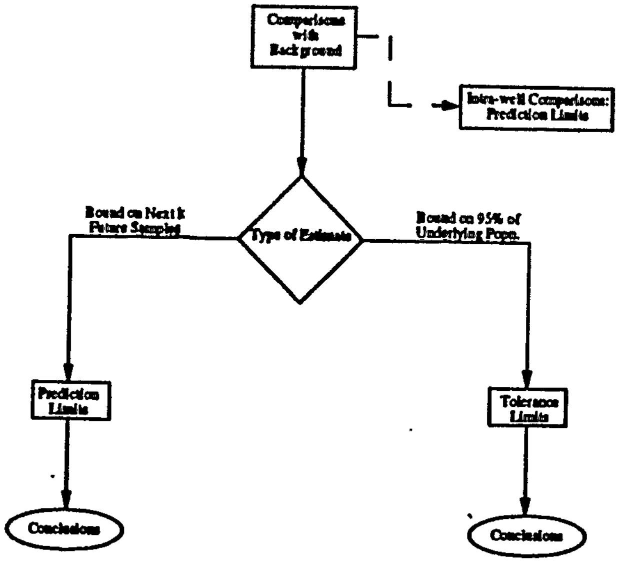
STATISTICAL INTERVALS



WHICH INTERVAL?

- Detection Monitoring: Background vs. Compliance
  - Unequal or unknown numbers of compliance samples per well: construct tolerance interval on background data
  - Known and equal numbers of compliance samples per well: construct prediction interval on background data
- Compliance Monitoring: Compliance Data vs. GWPS
  - CI containing upper 95th percentile for MCL or risk-based ACL
  - CI containing mean for background-determined ACL
- Intrawell Comparisons: single compliance well
  - Prediction interval on past data

COMPARISONS WITH BACKGROUND



Comparisons with Background

Intra-well Comparisons: Prediction Limits

Type of Estimate

Bound on Next k Future Sampling

Bound on 95% of Underlying Popn.

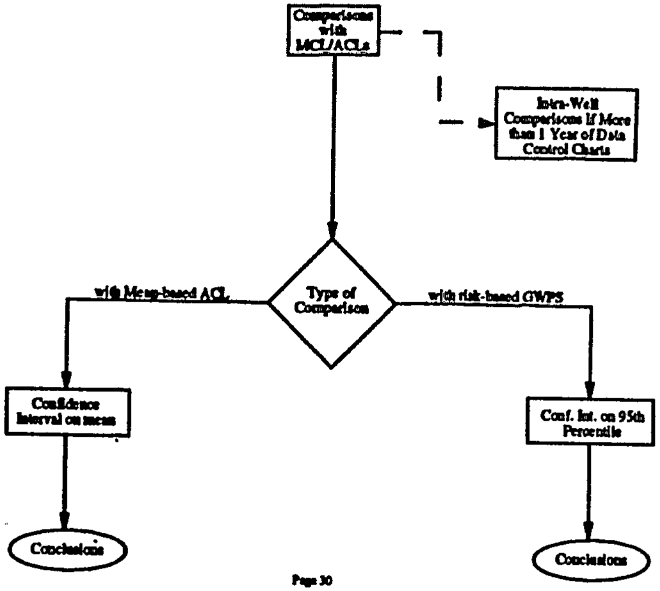
Prediction Limits

Tolerance Limits

Conclusions

Conclusions

COMPARISONS WITH MCL/ACLs



Comparisons with MCL/ACLs

Intra-Well Comparisons If More than 1 Year of Data Control Charts

Type of Comparison

with Mean-based ACL

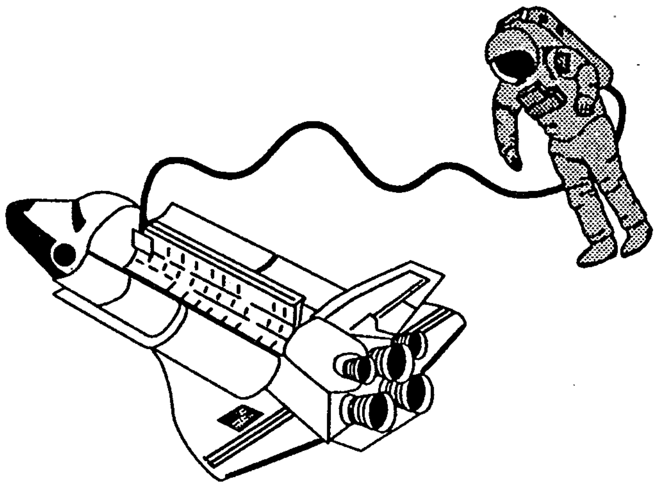
with risk-based GWPS

Confidence Interval on mean

Conf. Int. on 95th Percentile

Conclusions

Conclusions

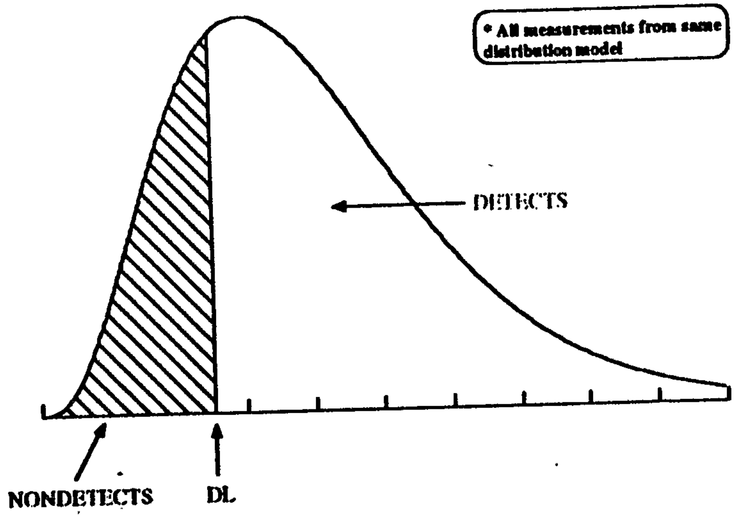


Section 9

HANDLING NON-DETECTS

ANOVA or t-test:	≤15% NDs ⇒	Substitute DL/2
	>15% NDs ⇒	Switch to non-parametric method (Wilcoxon or Kruskal-Wallis)
Interval Tests:	≤15% NDs ⇒	Can substitute DL/2, but see next option
	≤50% NDs ⇒	Cohen's or Aitchison's adjustment
	>50% NDs ⇒	Non-parametric test or interval
	≥90% NDs ⇒	Non-parametric test or interval or try Poisson-based prediction or tolerance limits
100% NDs:	No statistical test needed!!	

COHEN'S ADJUSTMENT


* All measurements from same distribution model

COHEN'S METHOD

$$\hat{\mu} = \bar{X}^{*} - \hat{\lambda}\left(\bar{X}^{*} - DL\right)$$

$$\hat{\sigma}^2 = \left(S^{*}\right)^2 + \hat{\lambda}\left(\bar{X}^{*} - DL\right)^2$$

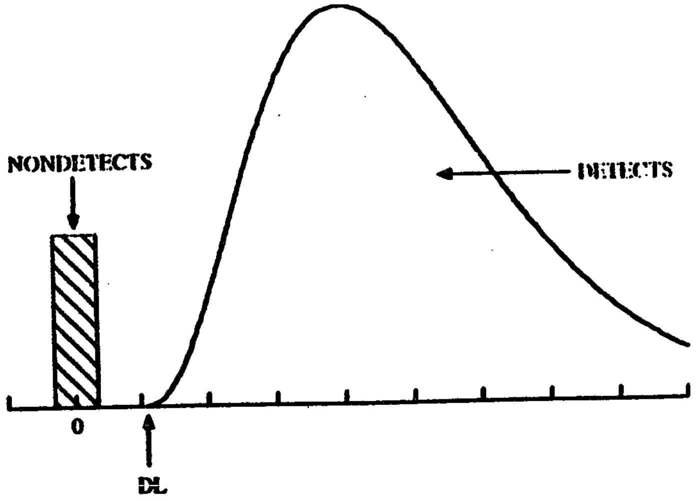
find $\hat{\lambda}$ in Table A-5 of Addendum

after calculating the percentage of non-detects and the parameter

$$\gamma = \frac{\left(S^{*}\right)^2}{\left(\bar{X}^{*} - DL\right)^2}$$

$\bar{X}^{*}$	=mean of detects
S^{*}	=SD of detects
DL	=detection limit

AITCHISON'S ADJUSTMENT



AITCHISON'S METHOD

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

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

$\bar{X}^*$ = mean of detects
S^* = SD of detects
d = # non-detects
n = total # samples

HOW TO DECIDE

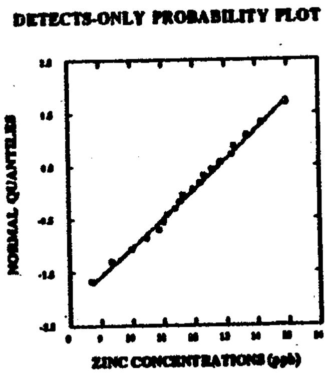
- Important to decide on appropriate model for censored data
- Consider physical aspects of facility and parameter being monitored
  - Is parameter naturally occurring?
- Comparing "censored" against "detects-only" probability plots may help

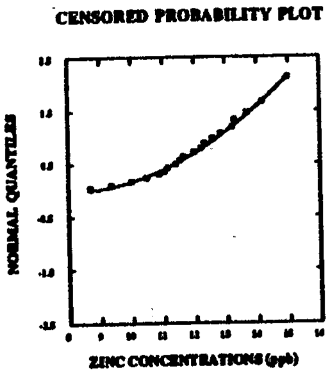
CENSORED VS. DETECTS-ONLY P-PLOTS

- Censored Probability Plot
  - Use to decide about Cohen's method
  - Construct plot as if nondetects were included, but only plot detects
  - If censored probability plot is linear, use Cohen's adjustment
- Detects-only Probability Plot
  - Use to decide about Aitchison's method
  - Construct plot after first excluding all nondetects
  - If plot is linear, use Aitchison's adjustment

CENSORED VS. DETECTS-ONLY PROBABILITY PLOTS

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





NON-PARAMETRIC TOLERANCE LIMITS

- Use if
  - Data are non-normal and non-lognormal
  - Data have frequent (>15%) non-detects
- In detection monitoring:
  - Upper TL=maximum of background data
  - Compare each compliance sample to this limit
- Requires large number of background samples to get decent coverage with 95% confidence

COPPER DATA

Background

Well 1	Well 2	Well 3
<5	9.2	<5
<5	<5	5.4
7.5	<5	6.7
<5	6.1	<5
<5	8.0	<5
<5	5.9	<5
6.4	<5	<5
6.0	<5	<5

Compliance Data

Well 4	Well 5
6.2	<5
<5	<5
7.8	5.6
10.4	<5

- Upper TL = 9.2
- One compliance well in violation
- 24 BG samples ⇒ minimum coverage = 88% with 95% confidence

NON-PARAMETRIC PREDICTION LIMITS

- Use with non-normal or frequently non-detect data
- Construction similar to non-parametric TL
  - Upper prediction limit (PL) = max of background in detection monitoring
  - Upper PL = max of past data in intrawell comparisons
- Interpretation different
  - Easier to predict that k future samples will fall below the known maximum
  - Fewer background data usually needed to achieve desired confidence level

POISSON PREDICTION LIMITS

- Goal: Upper Limit for sum of next k future samples
- If sum exceeds upper PL, test is triggered
- Calculate:

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

k= # future samples; n = #background samples

Tn= sum of background; z= normal distn. percentile

$$C = k/n$$

BENZENE

BW-1	BW-2	BW-3	BW-4	BW-5	CW
< 5	< 5	8.0	< 5	< 5	
< 5	< 5	10.6	< 5	< 5	
< 5	< 5	< 5	< 5	7.0	
< 5	< 5	< 5	< 5	< 5	< 5
< 5	< 5	12.0	< 5	< 5	9.5
< 5	< 5	< 5	< 5	< 5	10.2

Background n = 30 %NDs = 87%

Predict next 3 compliance samples $\Rightarrow k = 3$

$$\Rightarrow c = \frac{3}{30} = \frac{1}{10}$$

POISSON PREDICTION LIMIT EXAMPLE

- Set each ND to $\frac{1}{2}$ DL or 2.5
- Poisson count of sum = 26(2.5)+(8.0+10.6+12.0+7.0) = 102.6

  $\Rightarrow T_n = 102.6$
- Find $z_{.01} = 2.3263$ for 99%PL

$$T_3^* = \frac{1}{10}(102.6) + \frac{(2.3263)^2}{2(10)} + \frac{2.3263}{10}\sqrt{(102.6)[1+10] + \frac{(2.3263)^2}{4}} = 18.35$$

Since sum of CW samples = 22.2, have evidence of violation

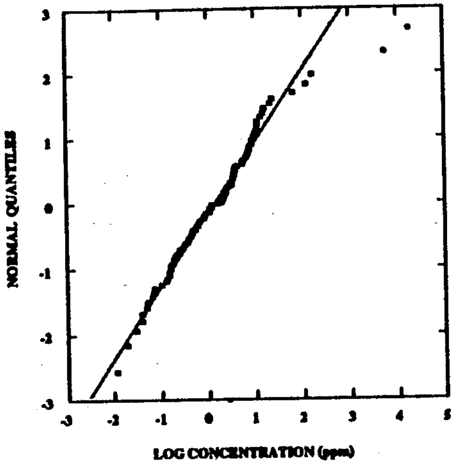
DETECTING OUTLIERS

- Definition A value that is very different from most other values (extreme value)
- Reasons
  - Contaminated equipment
  - Inconsistent sampling or analytical methodology
  - Data errors
  - True but extreme measurements
- What to Do
  - Correct value if you can
  - If error can be documented but correct value cannot be recovered ⇒ delete value
  - If no error ⇒ keep value

TESTING FOR OUTLIER(S)

- First look at probability plot of data excluding suspected outlier
  - Judge whether data are more normal or lognormal
  - If normal, run test on original data
  - If lognormal, run test on logged data
- Basic algorithm
  - List data in order
  - Calculate mean and SD of <u>all</u> data
  - Calculate T_n = (largest value - mean)/SD

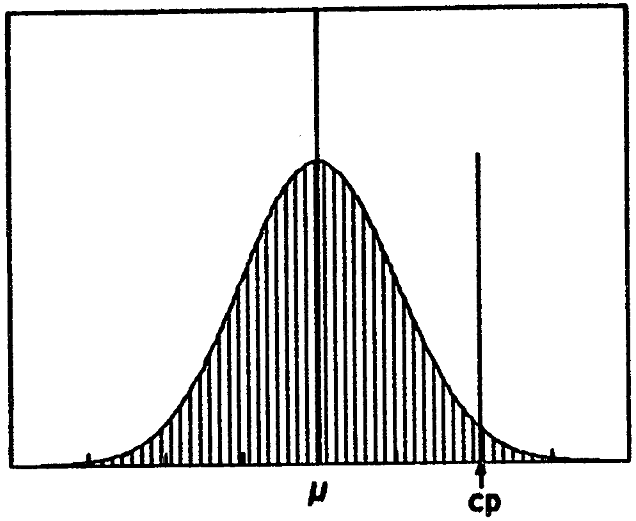
DETECTING OUTLIERS



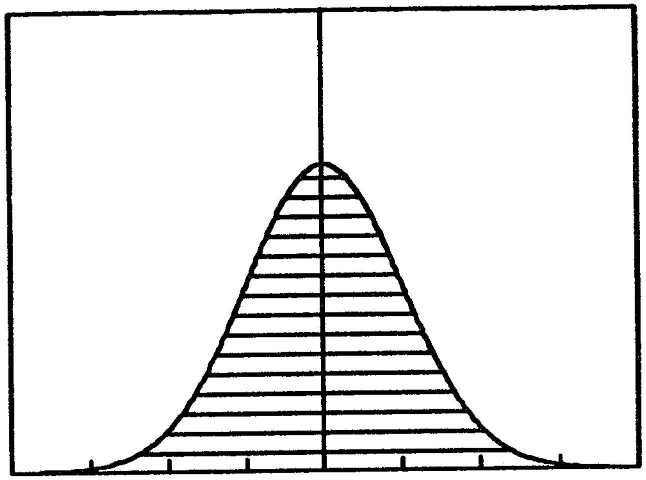
TESTING FOR OUTLIER(S) (cont.)

- Compare T_n to tabulated value
  - If T_n exceeds tabulated value, have evidence of a statistical outlier
  - Even if value is a statistical outlier, make sure reason can be documented

Section 10

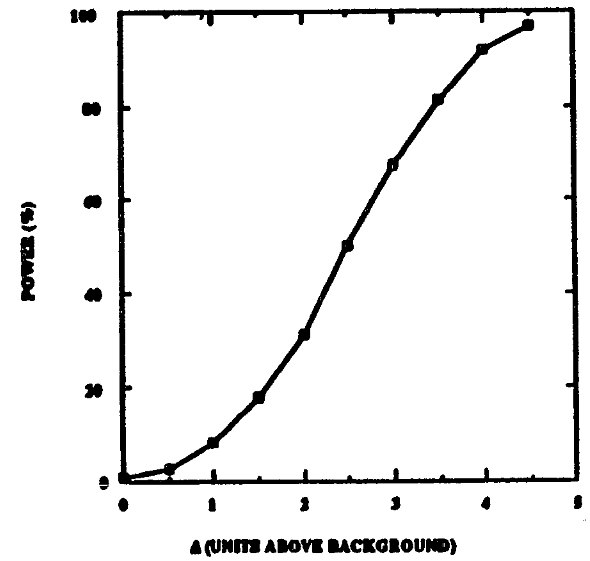


Page 1



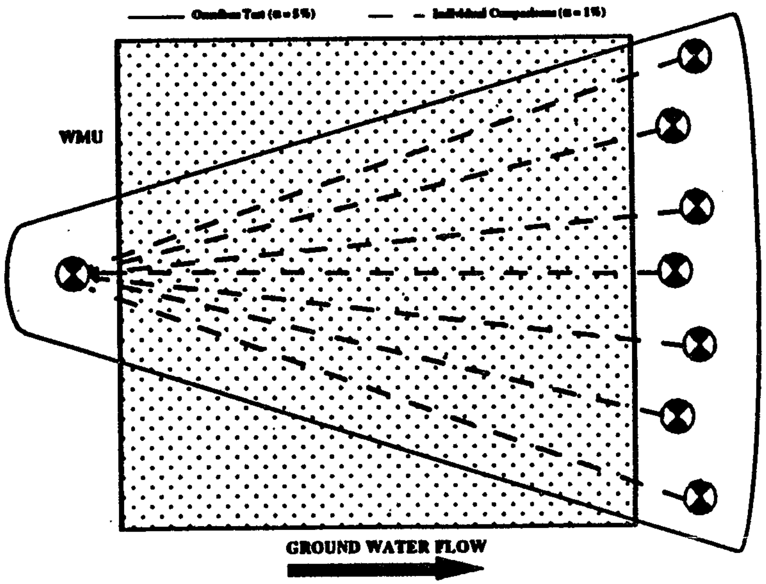
Page 2

EXAMPLE POWER CURVE



Page 3

MULTIPLE COMPARISONS



Page 4

EXPERIMENTWISE ERROR RATES

- General Formula
  - $\Pr\{\geq 1 \text{ false}(+)\} = 1-(1-\alpha)^{w \cdot c}$

    w = # wells

    c = # constituents
- For 100 Combinations
  - Expect average of 5 false (+)'s
  - $\Pr\{\geq 1 \text{ false}(+)\} = 1-(.95)^{100} = 99.4\%$

ERROR RATES: DEFINITIONS

- Comparisonwise Error
  - False positive rate of any single comparison
- Experimentwise Error
  - False positive rate of entire network of comparisons
  - Probability of false positive at one or more wells in network

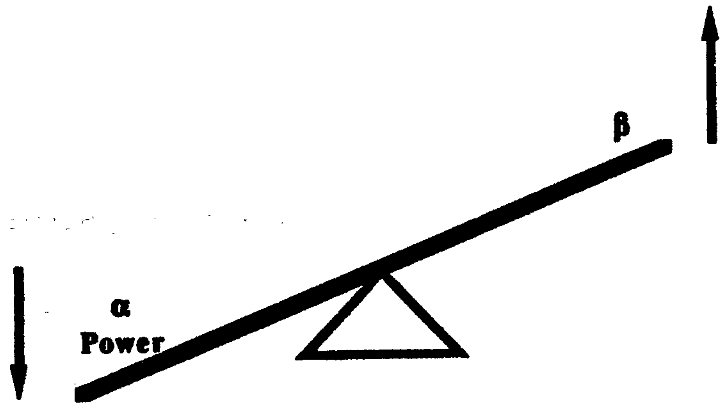
POSSIBLE STRATEGIES

- Bonferroni Approach
- ANOVA
- Retesting Individual Wells

BONFERRONI APPROACH

- Adjust significance level α for number of comparisons
  - Lowering α leads to more stringent individual tests
  - Less chance of individual false positives
  - Ultimate result: lower experimentwise error
- Example
  - Run 5 comparisons, each at $\alpha = 5\%$
  - Adjust comparisonwise error rate to $\alpha^* = \frac{\alpha}{5} = 1\%$
  - Experimentwise error rate drops from 22% to 5%

TYPE I vs. TYPE II ERROR



GOALS OF RETESTING

1. Keep facility-wide error rate low (~5%)

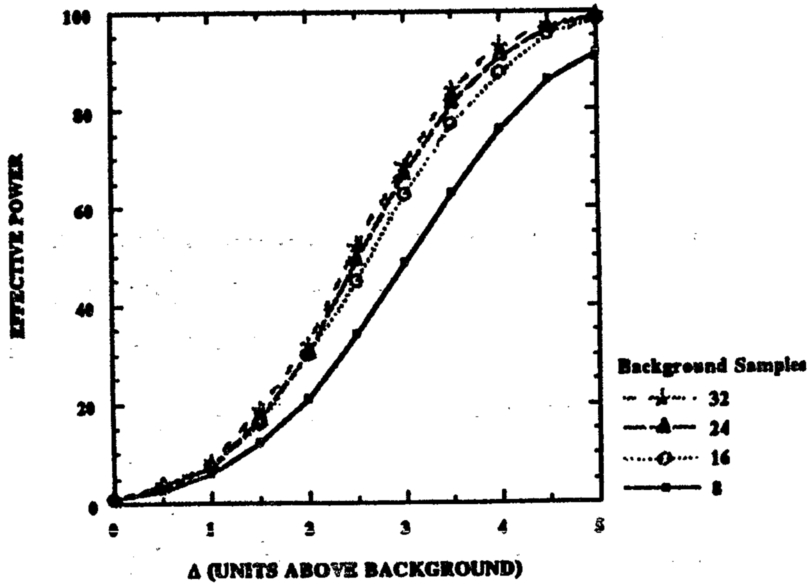
2. Keep effective power comparable to EPA Reference Power Curve

DEFINITIONS

A) Effective power: power of testing strategy to detect contamination at single well ("needle in a haystack" hypothesis)

B) EPA Reference Power Curve: power curve of a 99% prediction limit applied to a single well

EPA REFERENCE POWER CURVES



PARAMETRIC RETESTING

1. Collect background data

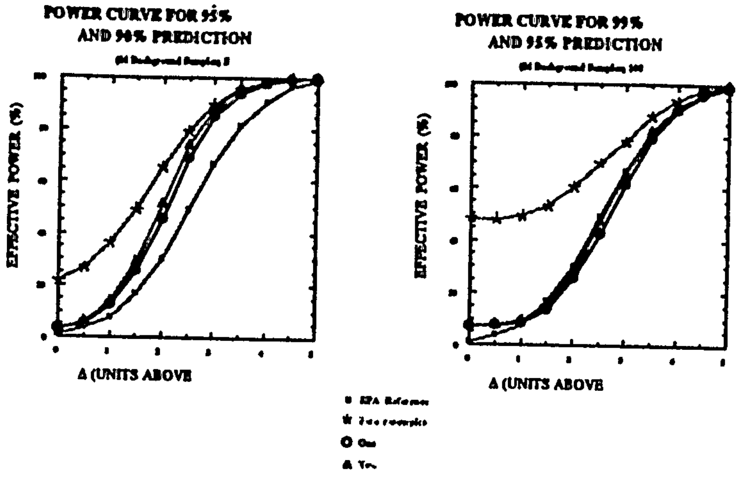
2. Construct 95% confidence upper tolerance limit on background

3. Compare new samples at compliance wells to upper tolerance limit

4. Resample any compliance wells that trigger the tolerance limit

5. Compare resamples to upper prediction limit derived from background data

6. Fail any well where one or more resamples flunks the prediction limit



POWER CURVE FOR 95% AND 90% PREDICTION

POWER CURVE FOR 99% AND 95% PREDICTION

NON-PARAMETRIC RETESTING

1. Collect background data

2. Construct non-parametric upper prediction limit on background data (find the maximum concentration)

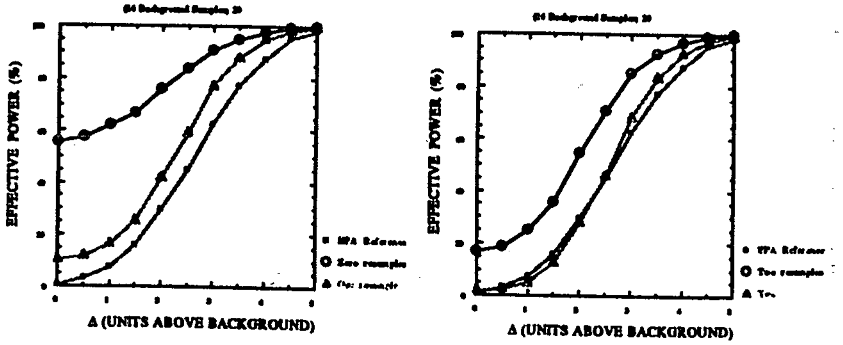
3. Compare new compliance well samples to upper prediction limit

4. Resample any compliance well that fails the upper prediction limit (may need one or two independent resamples)

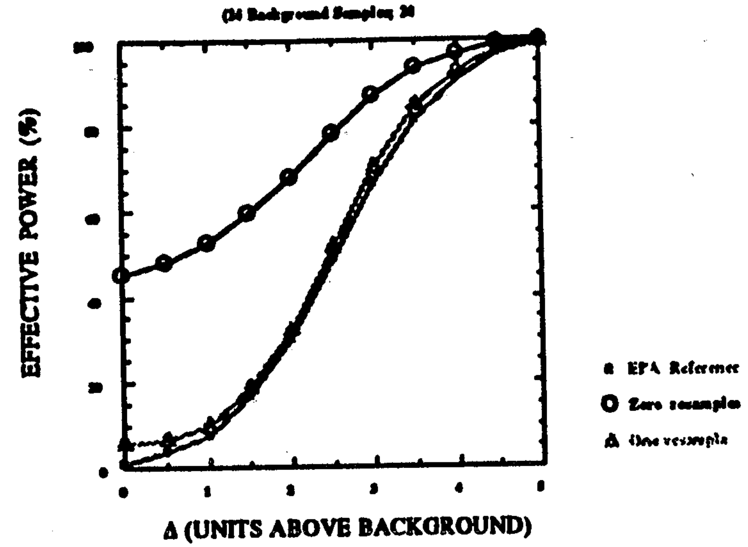
5. Compare resamples to prediction limit

6. Fail any compliance well for which one or more resamples fails the upper prediction limit

POWER CURVES FOR NON-PARAMETRIC PREDICTION LIMITS



POWER CURVE FOR NON-PARAMETRIC PREDICTION LIMITS



Notes