PART I. STATISTICAL DESIGN AND PHILOSOPHY

Chapter 1 provides introductory information, including the purposes and goals of the guidance, as well as its potential applicability to other environmental programs. **Chapter 2** presents a brief discussion of the existing regulations and identifies key portions of these rules which need to be addressed from a statistical standpoint, as well as some recommendations. In **Chapter 3**, fundamental statistical principles are highlighted which play a prominent role in the Unified Guidance including the notions of individual test false positive and negative decision errors and the accumulation of such errors across multiple tests or comparisons. **Chapter 4** sets the groundwater monitoring program context, the nature of formal statistical tests for groundwater and some caveats in identifying statistically significant increases. Typical groundwater monitoring scenarios also are described in this chapter. **Chapter 5** describes how to establish background and how to periodically update it. **Chapters 6** and **7** outline various factors to be considered when designing a reasonable statistical strategy for use in detection monitoring, compliance/assessment monitoring, or corrective action. Finally, **Chapter 8** summarizes the recommended statistical tests and methods, along with a concise review of assumptions, conditions of use, and limitations.

This page intentionally left blank

CHAPTER 1. OBJECTIVES AND POTENTIAL USE OF THIS GUIDANCE

1.1 OBJECTIVES

The fundamental goals of the RCRA groundwater monitoring regulations are fairly straightforward. Regulated parties are to accurately characterize existing groundwater quality at their facility, assess whether a hazardous constituent release has occurred and, if so, determine whether measured levels meet the compliance standards. Using accepted statistical testing, evaluation of groundwater quality should have a high probability of leading to correct decisions about a facility's regulatory status.

To implement these goals, EPA first promulgated regulations in 1980 (for interim status facilities) and 1982 (permitted facilities) for detecting contamination of groundwater at hazardous waste Subtitle C land disposal facilities. In 1988, EPA revised portions of those regulations found at 40 CFR Part 264, Subpart F. A similar set of regulations applying to Subtitle D municipal and industrial waste facilities was adopted in 1991 under 40 CFR Part 258. In April 2006, certain modifications were made to the 40 CFR Part 264 groundwater monitoring regulations affecting statistical testing and decision-making.

EPA released the ***Interim Final Guidance*** [IFG] in 1989 for implementing the statistical methods and sampling procedures identified in the 1988 rule. A second guidance document followed in July 1992 called ***Addendum to Interim Final Guidance*** [Addendum], which expanded certain techniques and also served as guidance for the newer Subpart D regulations.

As the RCRA groundwater monitoring program has matured, it became apparent that the existing guidance needed to be updated to adequately cover statistical methods and issues important to detecting changes in groundwater.[1] Research conducted in the area of groundwater statistics since 1992 has provided a number of improved statistical techniques. At the same time, experience gained in applying the regulatory statistical tests in groundwater monitoring contexts has identified certain constraints. Both needed to be factored into the guidance. This Unified Guidance document addresses these concerns and supercedes both the earlier IFG and Addendum.

The Unified Guidance offers guidance to owners and operators, EPA Regional and State personnel, and other interested parties in selecting, using, and interpreting appropriate statistical methods for evaluating data under the RCRA groundwater monitoring regulations. The guidance

[1] Some recommendations in EPA's **Statistical Training Course on Groundwater Monitoring** were developed to better reflect the reality of groundwater conditions at many sites, but were not generally available in published form. See RCRA Docket # EPA\530-R-93-003, 1993

identifies recent approaches and recommends a consistent framework for applying these methods. One key aspect of the Unified Guidance is providing a systematic application of the basic statistical principle of balancing false positives and negative errors in designing good testing procedures (*i.e.*, minimizing both the risk of falsely declaring a site to be out-of-compliance and of missing real evidence of an adverse change in the groundwater). Topics addressed in the guidance include basic statistical concepts, sampling design and sample sizes, selection of appropriate statistical approaches, how to check data and run statistical tests, and the interpretation of results. References for the suggested procedures and to more general statistical texts are provided. The guidance notes when expert statistical consultation may be advisable. Such guidance may also have applicability to other remedial activities as well.

Enough commonality exists in sampling, analysis, and evaluation under the RCRA regulatory requirements that the Unified Guidance often suggests relatively general strategies. At the same time, there may be situations where site-specific considerations for sampling and statistical analysis are appropriate or needed. EPA policy has been to promulgate regulations that are specific enough to implement, yet flexible in accommodating a wide variety of site-specific environmental factors. Usually this is accomplished by specifying criteria appropriate for the majority of monitoring situations, while at the same time allowing alternatives that are also protective of human health and the environment.

40 CFR Parts 264 and 258 allow the use of other sampling procedures and test methods[2] beyond those explicitly identified in the regulations,[3] subject to approval by the Regional Administrator or state Director. Alternative test methods must be able to meet the performance standards at §264.97(i) or §258.53(h). While these performance standards are occasionally specific, they are much less so in other instances. Accordingly, further guidance is provided concerning the types of procedures that should generally satisfy such performance standards.

Although the Part 264 and 258 regulations explicitly identify five basic formal statistical procedures for testing two- or multiple-sample comparisons characteristic of detection monitoring, the rules are silent on specific tests under compliance or corrective action monitoring when a groundwater protection standard is fixed (a one-sample comparison). The rules also require consideration of data patterns (normality, independence, outliers, non-detects, spatial and temporal dependence), but do not identify specific tests. This document expands the potential statistical procedures to cover these situations identified in earlier guidance, thus providing a comprehensive single EPA reference on statistical methods generally recommended for RCRA groundwater monitoring programs. Not every technique will be appropriate in a given situation, and in many cases more than one statistical approach can be used. The Unified Guidance is meant to be broad enough in scope to cover a high percentage of the potential situations a user might encounter.

The Unified Guidance is not designed as a treatise for statisticians; rather it is aimed at the informed groundwater professional with a limited background in statistics. Most methods discussed are well-known to statisticians, but not necessarily to regulators, groundwater engineers or scientists. A key thrust of the Unified Guidance has been to tailor the standard statistical techniques to the RCRA groundwater arena and its unique constraints. Because of this emphasis, not every variation of each test

[2] For example, §264.97(g)(2), §264.97(h)(5) and §258.53(g)(5)

[3] §264.97(g)(1), §264.97(h)(1-4), and §258.53(g)(1-4) respectively

is discussed in detail. For example, groundwater monitoring in a detection monitoring program is generally concerned with *increases* rather than *decreases* in concentration levels of monitored parameters. Thus, most detection monitoring tests in the Unified Guidance are presented as one-sided upper-tailed tests. In the sections covering compliance and corrective action monitoring (**Chapters 21** and **22** in **Part IV**), either one-sided lower-tail or upper-tail tests are recommended depending on the monitoring program. Users requiring two-tailed tests or additional information may need to consult other guidance or the statistical references listed at the end of the Unified Guidance.

The Unified Guidance is not intended to cover all statistical methods that might be applicable to groundwater. The technical literature is even more extensive, including other published frameworks for developing statistical programs at RCRA facilities. Certain statistical methods and general strategies described in the Unified Guidance are outlined in American Society for Testing and Materials [ASTM] documents entitled ***Standard Guide for Developing Appropriate Statistical Approaches for Groundwater Detection Monitoring Programs*** **(D6312-98[2005])** (ASTM, 2005) and ***Standard Guide for Applying Statistical Methods for Assessment and Corrective Action Environmental Monitoring Programs*** **(D7048-04)** (ASTM, 2004).

The first of these ASTM guidelines primarily covers strategies for detection monitoring, emphasizing the use of prediction limits and control charts. It also contains a series of flow diagrams aimed at guiding the user to an appropriate statistical approach. The second guideline covers statistical strategies useful in compliance/assessment monitoring and corrective action. While not identical to those described in the Unified Guidance, the ASTM guidelines do provide an alternative framework for developing statistical programs at RCRA facilities and are worthy of careful consideration.

EPA's primary consideration in developing the Unified Guidance was to select methods both consistent with the RCRA regulations, as well as straightforward to implement. We believe the methods in the guidance are not only effective, but also understandable and easy to use.

1.2 APPLICABILITY TO OTHER ENVIRONMENTAL PROGRAMS

The Unified Guidance is tailored to the context of the RCRA groundwater monitoring regulations. Some of the techniques described are unique to this guidance. Certain regulatory constraints and the nature of groundwater monitoring limit how statistical procedures are likely to be applied. These include typically small sample sizes during a given evaluation period, a minimum of annual monitoring and evaluation and typically at least semi-annual, often a large number of potential monitoring constituents, background-to-downgradient well comparisons, and a limited set of identified statistical methods. There are also unique regulatory performance constraints such as §264.97(i)(2), which requires a minimum single test false positive α level of 0.01 and a minimum 0.05 level for multiple comparison procedures such as analysis of variance [ANOVA].

There are enough commonalities with other regulatory groundwater monitoring programs (*e.g.*, certain distributional features of routinely monitored background groundwater constituents) to allow for more general use of the tests and methods in the Unified Guidance. Many of these test methods and the consideration of false positive and negative errors in site design are directly applicable to corrective action evaluations of solid waste management units under 40 CFR 264.101 and Comprehensive

Environmental Response, Compensation, and Liability Act [CERCLA] groundwater monitoring programs.

There are also comparable situations involving other environmental media to which the Unified Guidance statistical methods might be applied. Groundwater detection monitoring involves either a comparison between different monitoring stations (***i.e.***, downgradient compliance wells ***vs.*** upgradient wells) or a contrast between past and present data within a given station (***i.e.***, intrawell comparisons). To the extent that an environmental monitoring station is essentially fixed in location (***e.g.***, air quality monitors, surface water stations) and measurements are made over time, the same statistical methods may be applicable.

The Unified Guidance also details methods to compare background data against measurements from regulatory compliance points. These procedures (***e.g.***, Welch's *t*-test, prediction limits with retesting, ***etc.***) are designed to contrast multiple groups of data. Many environmental problems involve similar comparisons, even if the groups of data are not collected at fixed monitoring stations (***e.g.***, as in soil sampling). Furthermore, the guidance describes diagnostic techniques for checking the assumptions underlying many statistical procedures. Testing of normality is ubiquitous in environmental statistical analysis. Also common are checks of statistical independence in time series data, the assumption of equal variances across different populations, and the need to identify outliers. The Unified Guidance addresses each of these topics, providing useful guidance and worked out examples.

Finally, the Unified Guidance discusses techniques for comparing datasets against fixed numerical standards (as in compliance monitoring or corrective action). Comparison of data against a fixed standard is encountered in many regulatory programs. The methods described in **Part IV** of the Unified Guidance could therefore have wider applicability, despite being tailored to the groundwater monitoring data context.

EPA recognizes that many guidance users will make use of either commercially available or proprietary statistical software in applying these statistical methods. Because of their wide range of diversity and coverage, the Unified Guidance does not evaluate software usage or applicability. Certain software is provided with the guidance. The guidance limits itself to describing the basic statistical principles underlying the application of the recommended tests.

CHAPTER 2. REGULATORY OVERVIEW

This chapter generally summarizes the RCRA groundwater monitoring regulations under 40 CFR Parts 264, 265 and 258 applicable to this guidance. A second section identifies the most critical regulatory statistical issues and how they are addressed by this guidance. Finally, recommendations regarding interim status facilities and certain statistical methods in the regulations are presented at the end of the chapter.

2.1 REGULATORY SUMMARY

Section 3004 of RCRA directs EPA to establish regulations applicable to owners and operators of facilities that treat, store, or dispose of hazardous waste as may be necessary to protect human health and the environment. Section 3005 provides for the implementation of these standards under permits issued to owners and operators by EPA or authorized States. These regulations are codified in 40 CFR Part 264. Section 3005 also provides that owners and operators of facilities in existence at the time of the regulatory or statutory requirement for a permit, who apply for and comply with applicable requirements, may operate until a permit determination is made. These facilities are commonly known as interim status facilities, which must comply with the standards promulgated in 40 CFR Part 265.

EPA first promulgated the groundwater monitoring regulations under Part 265 for interim status surface impoundments, landfills and land treatment units ("regulated units") in 1980.[1] Intended as a temporary system for units awaiting full permit requirements, the rules set out a minimal detection and assessment monitoring system consisting of at least a single upgradient and three downgradient wells. Following collection of the minimum number of samples prescribed in the rule for four indicator parameters — pH, specific conductance, total organic carbon (TOC) and total organic halides (TOX) — and certain constituents defining overall groundwater quality, the owner/operator of a land disposal facility is required to implement a ***detection monitoring*** program. Detection monitoring consists of upgradient-to-downgradient comparisons using the Student's *t*-test of the four indicator parameters at no less than a .01 level of significance (α). The regulations refer to the use of "replicate" samples for contaminant indicator comparisons. Upon failure of a single detection-level test, as well as a repeated

[1] [45 FR 33232ff, May 19, 1980] Interim status regulations; later amended in 1983 and 1985

follow-up test, the facility is required to conduct an ***assessment*** program identifying concentrations of hazardous waste constituents from the unit in groundwater. A facility can return to detection monitoring if none of the latter constituents are detected. These regulations are still in effect today.

Building on the interim status rules, Subtitle C regulations for Part 264 permitted hazardous waste facilities followed in 1982,[2] where the basic elements of the present RCRA groundwater monitoring program are defined. In §264.91, three monitoring programs — ***detection monitoring***, ***compliance monitoring***, and ***corrective action*** — serve to protect groundwater from releases of hazardous waste constituents at certain regulated land disposal units (surface impoundments, waste piles, landfills, and land treatment). In developing permits, the Regional Administrator/State Director establishes groundwater protection standards [GWPS] under §264.92 using concentration limits [§264.94] for certain monitoring constituents [§264.93]. Compliance well monitoring locations are specified in the permit following the rules in §264.95 for the required compliance period [§264.96]. General monitoring requirements were established in §264.97, along with specific detection [§264.98], compliance [§264.99], and corrective action [§264.100] monitoring requirements. Facility owners and operators are required to sample groundwater at specified intervals and to use a statistical procedure to determine whether or not hazardous wastes or constituents from the facility are contaminating the groundwater.

As found in §264.91, detection monitoring is the first stage of monitoring when no or minimal releases have been identified, designed to allow identification of significant changes in the groundwater when compared to background or established baseline levels. Downgradient well observations are tested against established background data, including measurements from upgradient wells. These are known as two- or multiple-sample tests.

If there is statistically significant evidence of a release of hazardous constituents [§264.91(a)(1) and (2)], the regulated unit must initiate compliance monitoring, with groundwater quality measurements compared to the groundwater protection standards [GWPS]. The owner/operator is required to conduct a more extensive Part 261 Appendix VIII (later Part 264 Appendix IX)[3] evaluation to determine if additional hazardous constituents must be added to the compliance monitoring list.

Compliance/assessment as well as corrective action monitoring differ from detection monitoring in that groundwater well data are tested against the groundwater protection standards [GWPS] as established in the permit. These may be fixed health-based standards such as Safe Drinking Water Act [SDWA] maximum concentration limits [MCLs], §264.94 Table 1 values, a value defined from background, or alternate-concentration limits as provided in §264.94(a). Statistically, these are considered single-sample tests against a fixed limit (a background limit can either be a single- or two-sample test depending on how the limit is defined). An exceedance occurs when a constituent level is shown to be significantly greater than the GWPS or compliance standard.

If a hazardous monitoring constituent under compliance monitoring statistically exceeds the GWPS at any compliance well, the facility is subject to corrective action and monitoring under §264.100. Following remedial action, a return to compliance consists of a statistical demonstration that

[2] [47 FR 32274ff, July 26, 1982] Permitting Requirements for Land Disposal Facilities

[3] [52 FR 25942, July 9, 1987] List (Phase I) of Hazardous Constituents for Groundwater Monitoring; Final Rule

the concentrations of all relevant hazardous constituents lie below their respective standards. Although the rules define a three-tiered approach, the Regional Administrator or State Director can assess available information at the time of permit development to identify which monitoring program is appropriate [§264.91(b)].

Noteworthy features of the 1982 rule included retaining use of the four Part 265 indicator parameters, but allowing for additional constituents in detection monitoring. The number of upgradient and downgradient wells was not specified; rather the requirement is to have a sufficient number of wells to characterize upgradient and downgradient water quality passing beneath a regulated unit. Formalizing the "replicate" approach in the 1980 rules and the use of Student's *t*-test, rules under §264.97 required the use of aliquot replicate samples, which involved analysis of at least four physical splits of a single volume of water. In addition, Cochran's Approximation to the Behrens-Fisher [CABF] Student's *t*-test was specified for detection monitoring at no less than a .01 level of significance (α). Background sampling was specified for a one-year period consisting of four quarterly samples (also using the aliquot approach). The rules allowed use of a repeated, follow-up test subsequent to failure of a detection monitoring test. A minimum of semi-annual sampling was required.

In response to a number of concerns with these regulations, EPA amended portions of the 40 CFR Part 264 Subpart F regulations including statistical methods and sampling procedures on October 11, 1988.[4] Modifications to the regulations included requiring (if necessary) that owners and/or operators more accurately characterize the hydrogeology and potential contaminants at the facility. The rule also identifies specific performance standards in the regulations that all the statistical methods and sampling procedures must meet (discussed in a following section). That is, it is intended that the statistical methods and sampling procedures meeting these performance standards defined in §264.97 have a low probability both of indicating contamination when it is not present (Type I error), and of failing to detect contamination that actually is present (Type II error). A facility owner and/or operator must demonstrate that a procedure is appropriate for the site-specific conditions at the facility, and ensure that it meets the performance standards. This demonstration applies to any of the statistical methods and sampling procedures outlined in the regulation as well as any alternate methods or procedures proposed by facility owners and/or operators.

In addition, the amendments removed the required use of the CABF Student's t-test, in favor of five different statistical methods deemed to be more appropriate for analyzing groundwater monitoring data (discussed in a following section). The CABF procedure is still retained in Part 264, Appendix IV, as an option, but there are no longer specific citations in the regulations for this test. These newer procedures offer greater flexibility in designing a groundwater statistical program appropriate to site-specific conditions. A sixth option allows the use of alternative statistical methods, subject to approval by the Regional Administrator. EPA also instituted new groundwater monitoring sampling requirements, primarily aimed at ensuring adequate statistical sample sizes for use in analysis of variance [ANOVA] procedures, but also allowing alternative sampling plans to be approved by the Regional Administrator. The requirements identify the need for statistically independent samples to be used during evaluation. The Agency further recognizes that the selection of appropriate hazardous

[4] [53 FR 39720, October 11, 1988] 40 CFR Part 264: Statistical Methods for Evaluating Groundwater Monitoring From Hazardous Waste Facilities; Final Rule

constituent monitoring parameters is an essential part of a reliable statistical evaluation. EPA addressed this issue in a 1987 Federal Register notice.[5]

§264.101 requirements for corrective action at non-regulated units were added in 1985 and later.[6] The Agency determined that since corrective action at non-regulated units would work under a different program, these units are not required to follow the detailed steps of Subpart F monitoring.

In 1991, EPA promulgated Subtitle D groundwater monitoring regulations for municipal solid waste landfills in 40 CFR Part 258.[7] These rules also incorporate a three-tiered groundwater monitoring strategy (detection monitoring, assessment monitoring, and corrective action), and describe statistical methods for determining whether background concentrations or the groundwater protection standards [GWPS] have been exceeded.

The statistical methods and related performance standards in 40 CFR Part 258 essentially mirror the requirements found as of 1988 at 40 CFR Part 264 Subpart F, with certain differences. Minimum sampling frequencies are different than in the Subtitle C regulations. The rules also specifically provide for the GWPS using either current MCLs or standardized risk-based limits as well as background concentrations. In addition, a specific list of hazardous constituent analytes is identified in 40 CFR Part 258, Appendix I for detection-level monitoring, including the use of unfiltered (total) trace elements.

The 1988 and 1991 rule amendments identify certain statistical methods and sampling procedures believed appropriate for evaluating groundwater monitoring data under a variety of situations. Initial guidance to implement these methods was released in 1989 as: ***Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities: Interim Final Guidance*** [IFG]. The IFG covered basic topics such as checking distributional assumptions, selecting one of the methods and sampling frequencies. Examples were provided for applying the recommended statistical procedures and interpreting the results. Two types of compliance tests were provided for comparison to the GWPS — mean/median confidence intervals and upper limit tolerance intervals.

Given additional interest from users of the comparable regulations adopted for Subtitle D solid waste facilities in 1991, and with experience gained in implementing various tests, EPA actively sought to improve existing groundwater statistical guidance. This culminated in a July 1992 publication of: ***Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities: Addendum to Interim Final Guidance*** [Addendum].

The 1992 Addendum included a chapter devoted to retesting strategies, as well as new guidance on several non-parametric techniques not covered within the IFG. These included the Wilcoxon rank-sum test, non-parametric tolerance intervals, and non-parametric prediction intervals. The Addendum also included a reference approach for evaluating statistical power to ensure that contamination could be adequately detected. The Addendum did not replace the IFG — the two documents contained overlapping material but were mostly intended to complement one another based on newer information

[5] [52 FR 25942, July 9, 1987] op. cit.

[6] [50 FR 28747, July 15, 1985] Amended in 1987, 1993, and 1998

[7] [56 FR 50978, October 9, 1991] 40 CFR Parts 257 & 258: Solid Waste Disposal Facility Criteria: Final Rule, especially Part 258 Subpart E Groundwater Monitoring and Corrective Action

and comments from statisticians and users of the guidance. However, the Addendum changed several recommendations within the IFG and replaced certain test methods first published in the IFG. The two documents provided contradictory guidance on several points, a concern addressed by this guidance.

More recently in April 2006, EPA promulgated further changes to certain 40 CFR Part 264 groundwater monitoring provisions as part of the Burden Reduction Initiative Rule.[8] A brief summary of the regulatory changes and the potential effects on existing RCRA groundwater monitoring programs is provided. Four items of specific interest are:

- Elimination of the requirements to sample four successive times per statistical evaluation under §264.98(d) and §264.99(f) in favor of more flexible, site-specific options as identified in §264.97(g)(1)&(2);
- Removal of the requirements in §264.98(g) and §264.99(g) to annually sample *all* monitoring wells for Part 264 Appendix IX constituents in favor of a specific subset of wells;
- Modifications of these provisions to allow for a specific subset of Part 264 Appendix IX constituents tailored to site needs; and
- A change in the resampling requirement in §264.98(g)(3) from "within a month" to a site-specific schedule.

These changes to the groundwater monitoring provisions require coordination between the regulatory agency and owner/operator with final approval by the agency. Since the regulatory changes are not issued under the 1984 Hazardous and Solid Waste Amendments [HSWA] to RCRA, authorized State RCRA program adoption of these rules is discretionary. States may choose to maintain more stringent requirements, particularly if already codified in existing regulations. Where EPA has direct implementation authority, the provisions would go into effect following promulgation.

The first provision reaffirms the flexible approach in the Unified Guidance for detection monitoring sampling frequencies and testing options. State RCRA programs using the four-successive sampling requirements can still continue to do so under §264.97(g)(1), but the rule now allows for alternate sampling frequencies under §264.97(g)(2) in both detection and compliance monitoring. The second and third provisions provide more site- and waste-specific options for Part 264 Appendix IX compliance monitoring. The final provision provides more flexibility when resampling these Appendix IX constituents.

Since portions of the earlier and the most recent rules are still operative, all are considered in the present Unified Guidance. The effort to create this guidance began in 1996, with a draft release in December 2004, a peer review in 2005, and a final version completed in 2009.

[8] [71 FR 16862-16915] April 4, 2006

2.2 SPECIFIC REGULATORY FEATURES AND STATISTICAL ISSUES

This section describes critical portions of the RCRA groundwater monitoring regulations which the present guidance addresses. The regulatory language is provided below in bold and italics.[9] A brief discussion of each issue is provided in statistical terms and how the Unified Guidance deals with it.

2.2.1 STATISTICAL METHODS IDENTIFIED UNDER §264.97(h) AND §258.53(g)

The owner or operator will specify one of the following statistical methods to be used in evaluating groundwater monitoring data for each hazardous constituent which, upon approval by the Regional Administrator, will be specified in the unit permit. The statistical test chosen shall be conducted separately for each hazardous constituent in each well…

1. ***A parametric analysis of variance (ANOVA) followed by multiple comparison procedures to identify statistically significant evidence of contamination. The method must include estimation and testing of the contrasts between each compliance well's mean and the background mean levels for each constituent.***

2. ***An analysis of variance (ANOVA) based on ranks followed by multiple comparison procedures to identify statistically significant evidence of contamination. The method must include estimation and testing of the contrasts between each compliance well's median and the background median levels for each constituent.***

3. ***A tolerance interval or prediction interval procedure in which an interval for each constituent is established from the distribution of the background data, and the level of each constituent in each compliance well is compared to the upper tolerance or prediction limit.***

4. ***A control chart approach that gives control limits for each constituent.***

5. ***Another statistical method submitted by the owner or operator and approved by the Regional Administrator.***

Part III of the Unified Guidance addresses these specific tests, as applied to a detection monitoring program. It is assumed that statistical testing will be conducted separately for each hazardous constituent in each monitoring well. The recommended non-parametric ANOVA method based on ranks is identified in this guidance as the Kruskal-Wallis test. ANOVA tests are discussed in **Chapter 17**. Tolerance interval and prediction limit tests are discussed separately in **Chapters 17** and **18**, with particular attention given to implementing prediction limits with retesting when conducting multiple comparisons in **Chapter 19**. The recommended type of control chart is the combined Shewhart-CUSUM control chart test, discussed in **Chapter 20**. Where a groundwater protection standard is based on background levels, application of these tests is discussed in **Part I**, **Chapter 7** and **Part IV**, **Chapter 22.**

[9] The following discussions somewhat condense the regulatory language for ease of presentation and understanding. Exact citations for regulatory text should be obtained from the most recent Title 40 Code of Federal Regulations.

If a groundwater protection standard involves a fixed limit, none of the listed statistical methods in these regulations directly apply. Consequently, a number of other single-sample tests for comparison with a fixed limit are recommended in **Part IV**. Certain statistical limitations encountered when using ANOVA and tolerance level tests in detection and compliance monitoring are also discussed in these chapters. Additional use of ANOVA tests for diagnostic identification of spatial variation or temporal effects is discussed in **Part II**, **Chapters 13** and **14**.

2.2.2 PERFORMANCE STANDARDS UNDER §264.97(i) AND §258.53(h)

Any statistical method chosen under §264.97(h) [or §258.53(g)] for specification in the unit permit shall comply with the following performance standards, as appropriate:

1. ***The statistical method used to evaluate ground-water monitoring data shall be appropriate for the distribution of chemical parameters or hazardous constituents. If the distribution of the chemical parameters or hazardous constituents is shown by the owner or operator to be inappropriate for a normal theory test, then the data should be transformed or a distribution-free test should be used. If the distributions for the constituents differ, more than one statistical method may be needed.***

2. ***If an individual well comparison procedure is used to compare an individual compliance well constituent concentration with background constituent concentrations or a groundwater protection standard, the test shall be done at a Type I error level no less than 0.01 for each testing period. If a multiple comparisons procedure is used, the Type I experiment-wise error rate for each testing period shall be no less than 0.05; however, the Type I error of no less than 0.01 for individual well comparisons must be maintained. This performance standard does not apply to control charts, tolerance intervals, or prediction intervals.***

3. ***If a control chart approach is used to evaluate groundwater monitoring data, the specific type of control chart and its associated parameter values shall be proposed by the owner or operator and approved by the Regional Administrator if he or she finds it to be protective of human health and the environment.***

4. ***If a tolerance interval or a prediction interval is used to evaluate groundwater monitoring data, the levels of confidence, and for tolerance intervals, the percentage of the population that the interval must contain, shall be proposed by the owner or operator and approved by the Regional Administrator if he or she finds it protective of human health and the environment. These parameters will be determined after considering the number of samples in the background data base, the data distribution, and the range of the concentration values for each constituent of concern.***

5. ***The statistical method shall account for data below the limit of detection with one or more procedures that are protective of human health and the environment. Any practical quantification limit (pql) approved by the Regional Administrator under §264.97(h) [or §258.53(g)] that is used in the statistical method shall be the lowest concentration level that can be reliably achieved within specified limits of precision and accuracy during routine laboratory operating conditions available to the facility.***

6. If necessary, the statistical method shall include procedures to control or correct for seasonal and spatial variability as well as temporal correlation in the data.

These performance standards pertain to both the listed tests as well as others (such as those recommended in **Part IV** of the guidance for comparison to fixed standards). Each of the performance standards is addressed in **Part I** of the guidance for designing statistical monitoring programs and in **Part II** of the guidance covering diagnostic testing.

The ***first*** performance standard considers distributional properties of sample data; procedures for evaluating normality, transformations to normality, or use of non-parametric (distribution-free) methods are found in **Chapter 10**. Since some statistical tests also require an assumption of equal variances across groups, **Chapter 11** provides the relevant diagnostic tests. Defining an appropriate distribution also requires consideration of possible outliers. **Chapter 12** discusses techniques useful in outlier identification.

The ***second*** performance standard identifies minimum false positive error rates required when conducting certain tests. "Individual well comparison procedures" cited in the regulations include various ANOVA-type tests, Student's *t*-tests, as well as one-sample compliance monitoring/corrective action tests against a fixed standard. Per the regulations, these significance level (α) constraints do not apply to the other listed statistical methods — control charts, tolerance intervals, or prediction intervals.

When comparing an individual compliance well against background, the probability of the test resulting in a false positive or Type I error should be no *less* than 1 in 100 (1%). EPA required a minimum Type I error level for a given test and fixed sample size because false positive and negative rates are inversely related. By limiting Type I error rates to 1%, EPA felt that the risk of incurring false positives would be sufficiently low, while providing sufficient statistical power (i.e., the test's ability to control the false negative rate, that is, the rate of missing or not detecting true changes in groundwater quality).

Though a procedure to test an individual well like the Student's *t*-test may be appropriate for the smallest of facilities, more extensive networks of groundwater monitoring wells and monitoring parameters will generally require a ***multiple comparisons*** procedure. The 1988 regulations recognized this need in specifying a one-way analysis of variance [ANOVA] procedure as the method of choice for replacing the CABF Student's *t*-test. The *F*-statistic in an analysis of variance [ANOVA] does indeed control the site-wide or ***experiment-wise*** error rate when evaluating multiple upgradient and downgradient wells, at least for a single constituent. Using this technique allowed the Type I experiment-wise error rate *for each constituent* to be controlled to about 5% for each testing period.

To maintain adequate statistical power, the regulations also mandate that the ANOVA procedure be run at a *minimum* 5% false positive rate per constituent. But when a full set of well-constituent combinations are considered (particularly large suites of detection monitoring analytes at numerous compliance wells), the site-wide false positive rate can be much greater than 5%. The one-way ANOVA is inherently an interwell technique, designed to simultaneously compare datasets from different well locations. Constituents with significant natural spatial variation are likely to trigger the ANOVA *F*-statistic even in the absence of real contamination, an issue discussed in **Chapter 13**.

Control charts, tolerance intervals, and prediction intervals provide alternate testing strategies for simultaneously controlling false positive rates while maintaining adequate power to detect contamination during detection monitoring. Although the rules do not require a minimum nominal false positive rate as specified in the **second** performance standard, use of tolerance or prediction intervals combined with a retesting strategy can result in sufficiently low experiment-wise Type I error rates and the ability to detect real contamination. **Chapters 17**, **18** and **20** consider how tolerance limits, control charts, and prediction limits can be designed to meet the **third** and **fourth** performance standards specific to these tests considering the number of samples in background, the data distribution, and the range of concentration values for each constituent of concern [COC]. **Chapters 19** and **20** on multiple comparison procedures using prediction limits or control charts identify how retesting can be used to enhance power and meet the specified false positive objectives.

The ***fifth*** performance standard requires statistical tests to account for non-detect data. **Chapter 15** provides some alternative approaches for either adjusting or modeling sample data in the presence of reported non-detects. Other chapters include modifications of standard tests to properly account for the non-detect portion of data sets.

The ***sixth*** performance standard requires consideration of spatial or temporal (including seasonal) variation in the data. Such patterns can have major statistical consequences and need to be carefully addressed. Most classical statistical tests in this guidance require assumptions of data independence and stationarity. Independence roughly means that observing a given sample measurement does not allow a precise prediction of other sample measurements. Drawing colored balls from an urn at random illustrates and fits this requirement; in groundwater, sample volumes are assumed to be drawn more or less at random from the population of possible same-sized volumes comprising the underlying aquifer. Stationarity assumes that the population being sampled has a constant mean and variance across time and space. Spatial or temporal variation in the well means and/or variances can negate these test assumptions. **Chapter 13** considers the use of ANOVA techniques to establish evidence of spatial variation. Modification of the statistical approach may be necessary in this case; in particular, background levels will need to be established at each compliance well for future comparisons (termed *intrawell* tests). Control chart, tolerance limit, and prediction limit tests can be designed for intrawell comparisons; these topics are considered in **Part III** of this guidance.

Temporal variation can occur for a number of reasons — seasonal fluctuations, autocorrelation, trends over time, *etc*. **Chapter 14** addresses these forms of temporal variation, along with recommended statistical procedures. In order to achieve stationarity and independence, sample data may need to be adjusted to remove trends or other forms of temporal dependence. In these cases, the ***residuals*** remaining after trend removal or other adjustments are used for formal testing purposes. Correlation among monitoring constituents within and between compliance wells can occur, a subject also treated in this chapter.

When evaluating statistical methods by these performance standards, it is important to recognize that the ability of a particular procedure to operate correctly in minimizing unnecessary false positives while detecting possible contamination depends on several factors. These include not only the choice of significance level and test hypotheses, but also the statistical test itself, data distributions, presence or absence of outliers and non-detects, the presence or absence of spatial and temporal variation, sampling requirements, number of samples and comparisons to be made, and frequency of sampling. Since all of these statistical factors interact to determine the procedure's effectiveness, ***any proposed statistical***

procedure needs to be evaluated in its entirety, not by individual components. **Part I, Chapter 5** discusses evaluation of potential background databases considering all of the performance criteria.

2.2.3 HYPOTHESIS TESTS IN DETECTION, COMPLIANCE/ASSESSMENT, AND CORRECTIVE ACTION MONITORING

The Part 264 Subpart F groundwater monitoring regulations do not specifically identify the test hypotheses to be used in detection monitoring (§264.98), compliance monitoring (§264.99), and corrective action (§264.100). The same is true for the parallel Part 258 regulations for detection monitoring (§258.54), assessment monitoring (§258.55), and assessment of corrective measures (§258.56), as well as for evaluating interim status indicator parameters (§265.93) or Appendix III constituents. However, the language of these regulations as well as accepted statistical principles allow for clear definitions of the appropriate test hypotheses. Two- or multiple-sample comparisons (background *vs*. downgradient well data) are usually involved in detection monitoring (the comparison could also be made against an ACL limit based on background data). Units under detection monitoring are initially presumed not to be contributing a release to the groundwater unless demonstrated otherwise. From a statistical testing standpoint, the population of downgradient well measurements is assumed to be equivalent to or no worse than those of the background population; typically this translates into an initial or null hypothesis that the downgradient population mean is equal to or less than the background population mean. Demonstration of a release is triggered when one or more well constituents indicate statistically significant levels above background.

Compliance and corrective action tests generally compare single sets of sample data to a fixed limit or a background standard. The language of §264.99 indicates that a significant increase above a GWPS will demonstrate the need for corrective action. Consequently, the null hypothesis is that the compliance population mean (or perhaps an upper percentile) is at or below a given standard. The statistical hypothesis is thus quite similar to that of detection monitoring. In contrast, once an exceedance has been established and §264.100 is triggered, the null hypothesis is that a site is contaminated unless demonstrated to be significantly below the GWPS. The same principles apply to Part 258 monitoring programs. In Part 265, the detection monitoring hypotheses apply to an evaluation of the contaminant indicator parameters. The general subject of hypothesis testing is discussed in **Chapter 3**, and specific statistical hypothesis formulations are found in **Parts III** and **IV** of this guidance.

2.2.4 SAMPLING FREQUENCY REQUIREMENTS

Each of the RCRA groundwater monitoring regulations defines somewhat different minimum sampling requirements. §264.97(g)(1) & (2) provides two main options:

1. ***Obtaining a sequence of at least four samples taken at an interval that ensures, to the greatest extent technically feasible, that a statistically independent sample is obtained, by reference to the uppermost aquifer effective porosity, hydraulic conductivity, and hydraulic gradient, and the fate and transport characteristics of potential contaminants; or***

2. ***An alternate sampling procedure proposed by the owner or operator and approved by the Regional Administrator if protective of human health and the environment.***

Additional regulatory language in detection [§264.98(d)] and compliance [§264.99(f)] monitoring reaffirms the first approach:

> ***[A] a sequence of at least four samples from each well (background and compliance wells) must be collected at least semi-annually during detection/compliance monitoring…***

Interim status sampling requirements under §265.92[c] read as follows:

> ***(1) For all monitoring wells, the owner or operator must establish initial background concentrations or values of all parameters specified in paragraph (b) of this section. He must do this quarterly for one year;***
>
> ***(2) For each of the indicator parameters specified in paragraph (b)(3) of this section, at least four replicate measurements must be obtained for each sample and the initial background arithmetic mean and variance must be determined by pooling the replicate measurements for the respective parameter concentrations or values in samples obtained from upgradient wells during the first year.***

The requirements under Subtitle D §258.54(b) are somewhat different:

> ***The monitoring frequency for all constituents listed in Appendix I to this part,... shall be at least semi-annual during the active life of the facility.... A minimum of four independent samples from each well (background and downgradient) must be collected and analyzed for the Appendix I constituents... during the first semi-annual event. At least one sample from each well (background and downgradient) must be collected and analyzed during subsequent semi-annual events...***

The 1980 and 1982 regulations required four analyses of essentially a single physical sample for certain constituents, i.e., the four contaminant indicator parameters. The need for statistically independent data was recognized in the 1988 revisions to Part 264 and in the Part 258 solid waste requirements. In the latter rules, only a minimum single sample is required in successive semi-annual sampling events. Individual Subtitle C programs have also made use of the provision in §264.97(g)(2) to allow for fewer than four samples collected during a given semi-annual period, while other State programs require the four successive sample measurements. As noted, by the recent changes in the April 2006 Burden Reduction Rule, the explicit requirements to obtain at least four samples during the next evaluation period under 40 CFR §264.98(d) and §264.99(f) have been removed, allowing more general flexibility under the §264.97(g) sampling options. Individual State RCRA programs should be consulted as to whether these recent rule changes may be applicable.

The requirements of Parts 264 and 258 were generally intended to provide sufficient data for ANOVA-type tests in detection monitoring. However, control chart, tolerance limit, and prediction limit tests can be applied with as few as one new sample per evaluation, once background data are established. The guidance provides maximum flexibility in offering a range of prediction limit options in **Chapter 18** in order to address these various sample size requirements. Although not discussed in detail, the same conclusions pertain to the use of control charts or tolerance limits.

The use of the term "replicate" in the Part 265 interim status regulations can be a significant problem, if interpreted to mean repeat analyses of splits (or aliquots) of a single physical sample. The

regulations indicate the need for statistical independence among sample data for testing purposes. This guidance discusses the technical statistical problems that arise if replicate (aliquot) sample data are used with the required Student's *t*-test in Part 265. Thus, the guidance recommends, if possible, that interim status statistical evaluations be based on independent sample data as discussed in **Chapters 13** and **14** and at the end of this chapter. A more standardized Welch's version of the Student-t test for unequal variances is provided as an alternative to the CABF Student's *t*-test.

2.2.5 GROUNDWATER PROTECTION STANDARDS

Part 265 does not use the term groundwater protection standards. A first-year requirement under §265.92(c)(1) is:

> ***For all monitoring wells, the owner or operator must establish background concentrations or values of all parameters specified in paragraph (b) of this section. He must do this quarterly for one year.***

Paragraph (b) includes water supply parameters listed in Part 265 Appendix III, which also provides a Maximum Level for each constituent. If a facility owner or operator does *not* develop and implement an assessment plan under §265.93(d)(4), there is a requirement in §265.94(a)(2) to report the following information to the Regional Administrator:

> ***(i) During the first year when initial background concentrations are being established for the facility: concentrations or values of the parameters listed in §265.92(b)(1) for each groundwater monitoring well within 15 days after completing each quarterly analysis. The owner or operator must separately identify for each monitoring well any parameters whose concentrations or value has been found to exceed the maximum contaminant levels in Appendix III.***

Since the Part 265 regulations are explicit in requiring a one-to-one comparison, no statistical evaluation is needed or possible.

§264.94(a) identifies the permissible concentration limits as a GWPS under §264.92:

> ***The Regional Administrator will specify in the facility permit concentrations limits in the groundwater for hazardous constituents established under §264.93. The concentration of a constituent:***
>
> ***(1) must not exceed the background level of that constituent in the groundwater at the time the limit is specified in the permit; or***
>
> ***(2) for any of the constituents listed in Table 1, must not exceed the respective value given in that table if the background level is below the value given in Table 1; or***
>
> ***(3) must not exceed an alternate limit established by the Regional Administrator under paragraph (b) of this section.***

The RCRA Subtitle D regulations establish the following standards under §258.55(h) and (i):

> ***(h) The owner or operator must establish a groundwater protection standard for each Appendix II constituent detected in groundwater. The groundwater protection standard shall be:***
>
> > ***(1) For constituents for which a maximum contaminant level (MCL) has been promulgated under Section 1412 of the Safe Drinking Water Act (codified) under 40 CFR Part 141, the MCL for that constituent;***
> >
> > ***(2) for constituents for which MCLs have not been promulgated, the background concentration for the constituent established from wells in accordance with §258.51(a)(1); or***
> >
> > ***(3) for constituents for which the background level is higher than the MCL identified under paragraph (h)(1) of this section or health based levels identified under §258(i)(1), the background concentration.***
>
> ***(i) The Director of an approved State program may establish an alternative groundwater protection standard for constituents for which MCLs have not been established. These groundwater protection standards shall be appropriate health based levels that satisfy the following criteria:***
>
> > ***(1) the level is derived in a manner consistent with Agency guidelines for assessing health risks or environmental pollutants [51 FR 33992, 34006, 34014, 34028, Sept. 24, 1986]***
> >
> > ***(2) to (4)... [other detailed requirements for health risk assessment procedures]***

The two principal alternatives for defining a groundwater protection standard [GWPS] are either a limit based on background data or a fixed health-based value (***e.g.***, MCLs, §264.94 Table 1 values, or a calculated risk limit). The Unified Guidance discusses these two basic kinds of standards in **Chapters 7 and 21**. If a background limit is applied, some definition of how the limit is constructed from prior sample data is required at the time of development. For fixed health-based limits, the regulatory program needs to consider the statistical characteristic of the data (***e.g.***, mean, median, upper percentile) that best represents the standard in order to conduct appropriate statistical comparisons. This subject is also discussed in **Chapter 21**; the guidance provides a number of testing options in this regard.

2.3 UNIFIED GUIDANCE RECOMMENDATIONS

2.3.1 INTERIM STATUS MONITORING

As discussed in **Chapter 14**, replicates required for the four contaminant indicator parameters are not statistically independent when analyzed as aliquots or splits from a single physical sample. This results in incorrect estimates of variance and the degrees of freedom when used in a Student's *t*-test. One of the most important revisions in the 1988 regulations was to require that successive samples be independent. Therefore, at a minimum, the Unified Guidance recommends that only **independent** water quality sample data be applied to the detection monitoring Student's *t*-tests in **Chapter 16**.

There are other considerations limiting the application of these tests as well. As noted in **Chapter 5**, at least two of the indicator parameters (pH and specific conductance) are likely to exhibit natural spatial differences among monitoring wells. Depending on site groundwater characteristics, TOC and TOX may also vary spatially. TOX analytical limitations described in SW-846[10] also note that levels of TOX are affected by inorganic chloride levels, which themselves can vary spatially by well. In short, all four indicator parameters may need to be evaluated on an intrawell basis, *i.e.*, using historical data from compliance monitoring wells.

Since this option is not identified in existing Part 265 regulations for indicator detection monitoring, a more appropriate strategy is to develop an alternative ***groundwater quality assessment monitoring plan*** under §265.90(d)(3) and (4) and §265.93(d)(3) and (4). These sections of the regulations require evaluation of hazardous waste constituents reasonably derived from the regulated unit (either those which served as a basis for listing in Part 265 Appendix VII or which are found in §261.24 Table 1). Interim status units subject to a permit are also subject to the groundwater contaminant information collection provisions under §270.14[c], which potentially include all hazardous constituents (a wider range of contaminants, *e.g.*, Part 264 Appendix IX) reasonably expected from the unit. While an interim status facility can return to indicator detection monitoring if no hazardous constituent releases have been identified, such a return is itself optional.

EPA recommends that interim status facilities develop the §265.90(d)(3) & (4) alternative groundwater quality assessment monitoring plan, if possible, using principles and procedures found in this guidance for monitoring design and statistical evaluation. Unlike Part 264 monitoring, there are no formal compliance/corrective action steps associated with statistical testing. A regulatory agency may take appropriate enforcement action if data indicate a release or significant adverse effect. The monitoring plan can be applied for an indefinite period until permit development. Multi-year collection of semi-annual or quarterly hazardous constituent data is more determinative of potential releases. The facility or the regulatory agency may also wish to continue evaluation of some or all of the Part 265 water quality indicators. Eventually these groundwater data can be used to establish which monitoring program(s) may be appropriate at the time of permit development under §264.91(b).

2.3.2 PARTS 264 AND 258 DETECTION MONITORING METHODS

As described in **Chapter 13**, many of the commonly monitored inorganic analytes exhibit natural spatial variation among wells. Since the two ANOVA techniques in §264.97(h) and §258.53(g) depend on an assumption of a single common background population, these tests may not be appropriate in many situations. Additionally, at least 50% of the data should be detectable in order to compare either well means or medians. For many hazardous trace elements, detectable percentages are considerably lower. Interwell ANOVA techniques would also not be generally useful in these cases. ANOVA may find limited applicability in detection monitoring with trace organic constituents, especially where downgradient levels are considerably higher than background and there is a high percentage of detects. Based on ranks alone, it may be possible to determine that compliance well(s) containing one or more hazardous constituents exceed background. However, the Unified Guidance recommends avoiding ANOVA techniques in the limiting situations just described.

[10] Test Methods for Evaluating Solid Waste (SW-846), EPA OSWER, 3rd Edition and subsequent revisions, Method 9020B, September 1994

Another detection monitoring method receiving less emphasis in this guidance is the tolerance limit. In previous guidance, an upper tolerance limit based on background was suggested to identify significant increases in downgradient well concentration levels. While still acceptable by regulation (*e.g.*, under existing RCRA permits), use of prediction limits are preferable to tolerance limits in detection monitoring for the following reasons. The construction of a tolerance limit is nearly identical to that of a prediction limit. In parametric normal distribution applications, both methods use the general formula: $\bar{x} + \kappa s$. The kappa (κ) multiplier varies depending on the coverage and confidence levels desired, but in both cases some multiple of the standard deviation (*s*) is added or subtracted from the sample mean ($\bar{x}$). For non-parametric limits, the similarity is even more apparent. Often the identical statistic (*e.g.*, the maximum observed value in background) can either be used as an upper prediction limit or an upper tolerance limit, with only a difference in statistical interpretation.

More fundamentally, given the wide variety of circumstances in which retesting strategies are now encouraged and even necessary, the mathematical underpinnings of retesting with *prediction limits* are well established while those for retesting with *tolerance limits* are not. Monte Carlo simulations were originally conducted for the 1992 Addendum to develop appropriate retesting strategies involving tolerance limits. Such simulations were found insufficient for the Unified Guidance.[11]

While the simultaneous prediction limits presented in the Unified Guidance consider the actual number of comparisons in defining exact false positive error rates, some tolerance limit approaches (including past guidance) utilized an approximate and less precise pre-selected low level of probability. On balance, there is little practical need for recommending two highly similar (but not identical) methods in the Unified Guidance, both for the reasons just provided and to avoid confusion of which method to use. The final regulation-specified detection monitoring method — *control charts* — is comparable to prediction limits, but possesses some unique benefits and so is also recommended in this guidance.

2.3.3 PARTS 264 AND 258 COMPLIANCE/ASSESSMENT MONITORING

A second use of tolerance limits recommended in earlier guidance was for comparing downgradient monitoring well data to a fixed limit during compliance/assessment monitoring. In this case, an upper tolerance limit constructed on each compliance well data set could be used to identify non-compliance with a fixed GWPS limit. Past guidance also used upper confidence limits around an upper proportion in defining these tolerance limits. A number of problems were identified using this approach.

A tolerance limit makes statistical sense if the limit represents an upper percentile, *i.e.*, when a limit is not to be exceeded by more than, for instance, 1% or 5% or 10% of future individual concentration values. However, GWPS limits can also be interpreted as long-term averages, *e.g.*, chronic risk-based values, which are better approximated by a statistic like the mean or median. **Chapters 7 &**

[11] 1) there were minor errors in the algorithms employed; 2) Davis and McNichols (1987) demonstrated how to compute exact kappa multipliers for prediction limits using a numerical algorithm instead of employing an inefficient simulation strategy; and 3) further research (as noted in **Chapter 19**) done in preparation of the guidance has shown that repeated prediction limits are more statistically powerful than retesting strategies using tolerance limits for detecting changes in groundwater quality.

22 discuss important considerations when identifying the appropriate statistical parameter to be compared against a fixed GWPS limit.

More importantly, since the upper confidence level of tolerance limit overestimates the true population proportion by design, demonstrating an exceedance of a GWPS by this limit does not necessarily indicate that the corresponding population proportion also exceeds the standard, leading to a high false positive rate. Therefore, the Unified Guidance recommends that the compliance/assessment monitoring null hypothesis be structured so that the compliance population characteristic (***e.g.***, mean, median, upper percentile) is assumed to be less than or equal to the fixed standard unless demonstrated otherwise. The correct test statistic in this situation is then the lower confidence limit. The upper confidence limit is used in corrective action to identify whether a constituent has returned to compliance.

To ensure consistency with the underlying statistical presumptions of compliance/assessment monitoring (see **Chapter 4**) and to maintain control of false positive rates, the Unified Guidance recommends that this tolerance interval approach be replaced with a more coherent and comprehensive strategy based on the use of confidence intervals (see **Chapters 21** and **22**). Confidence intervals can be applied in a consistent fashion to GWPS concentration limits representing either long-term averages or upper percentiles.

CHAPTER 3. KEY STATISTICAL CONCEPTS

The success of any discipline rests on its ability to accurately model and explain real problems. Spectacular successes have been registered during the past four centuries by the field of mathematics in modeling fundamental processes in mechanics and physics. The last century, in turn, saw the rise of statistics and its fundamental theory of ***estimation*** and ***hypothesis testing***. All of the tests described in the Unified Guidance are based upon this theory and involve the same key concepts. The purpose of this chapter is to summarize the statistical concepts underlying the methods presented in the Unified Guidance, and to consider each in the practical context of groundwater monitoring. These include:

- Statistical inference: the difference between samples and populations; the concept of sampling.
- Common statistical assumptions used in groundwater monitoring: statistical independence, stationarity, lack of outliers, and normality.
- Frequently-used statistical measures: mean, standard deviation, percentiles, correlation coefficient, coefficient of variation, ***etc.***
- Hypothesis testing: How probability distributions are used to model the behavior of groundwater concentrations and how the statistical evidence is used to "prove" or "disprove" the validity of competing models.
- Errors in hypothesis testing: What false positives (Type I errors) and false negatives (Type II errors) really represent.
- Sampling distributions and the Central Limit Theorem: How the statistical behavior of test statistics differs from that of individual population measurements.
- Statistical power and power curves: How the ability to detect real contamination depends on the size or degree of the concentration increase.
- Type I ***vs.*** Type II errors: The tradeoff between false positives and false negatives; why it is generally impossible to minimize both kinds of error simultaneously.

3.1 INTRODUCTION TO GROUNDWATER STATISTICS

This section briefly covers some basic statistical terms and principles used in this guidance. All of these topics are more thoroughly discussed in standard textbooks. It is presumed that the user already has some familiarity with the following terms and discussions.

Statistics is a branch of applied mathematics, dealing with the description, understanding, and modeling of data. An integral part of statistical analysis is the testing of competing mathematical models and the management of data uncertainty. Uncertainty is present because measurement data exhibit *variability*, with limited knowledge of the medium being sampled. The fundamental aim of almost every statistical analysis is to draw *inferences*. The data analyst must *infer* from the observed data something about the physical world without knowing or seeing all the possible facts or evidence. So the question becomes: how closely do the measured data mimic reality, or put another way, to what extent do the data correctly identify a physical truth (*e.g.*, the compliance well is contaminated with arsenic above regulatory limits)?

One way to ascertain whether an aquifer is contaminated with certain chemicals would be to exhaustively sample and measure every physical volume of groundwater underlying the site of interest. Such a collection of measurements would be impossible to procure in practice and would be infinite in size, since sampling would have to be continuously conducted over time at a huge number of wells and sampling depths. However, one would possess the entire *population* of possible measurements at that site and the exact statistical *distribution* of the measured concentration values.

A statistical *distribution* is an organized summary of a set of data values, sorted into the relative frequencies of occurrence of different measurement levels (*e.g.*, concentrations of 5 ppb or less occur among 30 percent of the values, or levels of 20 ppb or more only occur 1 percent of the time). More generally, a distribution may refer to a mathematical model (known as a *probability distribution*) used to represent the shape and statistical characteristics of a given population and chosen according to one's experience with the type of data involved.

By contrast to the population, a statistical *sample* is a finite subset of the population, typically called a *data set*. Note that the statistical definition of sample is usually different from a geological or hydrological definition of the same term. Instead of a physical volume or mass, a statistical sample is a collection of measurements, *i.e.*, a set of numbers. This collection might contain only a single value, but more generally has a number of measurements denoted as the *sample size*, *n*.

Because a sample is only a partial representation of the population, an inference is usually desired in order to conclude something from the observed data about the underlying population. One or more numerical characteristics of the population might be of interest, such as the true *average* contaminant level or the *upper 95th percentile* of the concentration distribution. Quantities computed from the sample data are known as *statistics*, and can be used to reasonably *estimate* the desired but unknown population characteristics. An example is when testing sample data against a regulatory standard such as a maximum concentration limit [MCL] or background level. *A mean sample estimate* of the average concentration can be used to judge whether the corresponding population characteristic — the true mean concentration (denoted by the Greek letter μ) — exceeds the MCL or background limit.

The accuracy of these estimates depends on how *representative* the sample measurements of the underlying population are. In a representative sample, the distribution of sample values have the best

chance of closely matching the population distribution. Unfortunately, the degree of representativeness of a given sample is almost never known. So it quite important to understand precisely how the sample values were obtained from the population and to explore whether or not they *appear* representative. Though there is no guarantee that a sample will be adequate, the best protection against an unrepresentative sample is to select measurements from the population *at random*. A *random sample* implies that each potential population value has an equivalent chance of being selected depending only on its likelihood of occurrence. Not only does random sampling guard against selection of an unrepresentative portion of the population distribution, it also enables a mathematical estimate to be drawn of the statistical uncertainty associated with the ability of a given sample to represent the desired characteristic of the population. It can be very difficult to gauge the uncertainty surrounding a sample collected haphazardly or by means of professional judgment.

As a simple example, consider an urn filled with red and green balls. By thoroughly mixing the urn and blindly sampling (*i.e.*, retrieving) 10 percent of the balls, a very nearly random sample of the population of balls will be obtained, allowing a fair estimate of the true overall proportion of one color or the other. On the other hand, if one looked into the urn while sampling and only picked red balls or tried to alternate between red and green, the sample would be far from random and likely unrepresentative of the true proportions.

At first glance, groundwater measurements obtained during routine monitoring would not seem to qualify as random samples. The well points are generally not placed in random locations or at random depths, and the physical samples are usually collected at regular, pre-specified intervals. Consequently, further distinctions and assumptions are necessary when performing statistical evaluations of groundwater data. First, the distribution of a given contaminant may not be *spatially uniform* or *homogeneous*. That is, the local distribution of measured values at one well may not be the same as at other wells. Because this is often true for naturally-occurring groundwater constituents, the statistical population(s) of interest may be well-specific. A statistical sample gathered from a particular well must then be treated as potentially representative only of that well's local population. On the other hand, samples drawn from a number of reference background wells for which no significant differences are indicated, may permit the pooled data to serve as an estimate of the overall well field behavior for that particular monitoring constituent.

The distribution of a contaminant may also not be *temporally uniform* or *stationary over time*. If concentration values indicates a trend, perhaps because a plume intensifies or dissipates or natural in-situ levels rise or fall due to drought conditions, *etc.*, the distribution is said to be *non-stationary*. In this situation, some of the measurements collected over time may not be representative of current conditions within the aquifer. Statistical adjustments might be needed or the data partitioned into usable and unusable values.

A similar difficulty is posed by *cyclical* or *seasonal* trends. A long-term constituent concentration average at a well location or the entire site may essentially be constant over time, yet temporarily fluctuate up and down on a seasonal basis. Given a fixed interval between sampling events, some of this fluctuation may go unobserved due to the non-random nature of the sampling times. This could result in a sample that is unrepresentative of the population *variance* and possibly of the population *mean* as well. In such settings, a shorter (*i.e.*, higher frequency) or staggered sampling interval may be needed to better capture key characteristics of the population as a part of the distribution of sample measurements.

The difficulties in identifying a valid statistical framework for groundwater monitoring highlight a fundamental assumption governing almost every statistical procedure and test. It is the presumption that sample data from a given population should be *independent* and *identically distributed*, commonly abbreviated as *i.i.d.* All of the mathematics and statistical formulas contained in this guidance are built on this basic assumption. If it is not satisfied, statistical conclusions and test results may be invalid or in error. The associated statistical uncertainty may be different than expected from a given test procedure.

Random sampling of a single, fixed, stationary population will guarantee independent, identically-distributed sample data. Routine groundwater sampling typically does not. Consequently, the Unified Guidance discusses both below and in later chapters what assumptions about the sample data must be routinely or periodically checked. Many but not all of these assumptions are a simple consequence of the *i.i.d.* presumption. The guidance also discusses how sampling ought to be conducted and designed to get as close as possible to the *i.i.d.* goal.

3.2 COMMON STATISTICAL ASSUMPTIONS

Every statistical test or procedure makes certain assumptions about the data used to compute the method. As noted above, many of these assumptions flow as a natural consequence of the presumption of *independent, identically-distributed* data (*i.i.d.*). The most common assumptions are briefly described below:

3.2.1 STATISTICAL INDEPENDENCE

A major advantage of truly random sampling of a population is that the measurements will be *statistically independent.* This means that observing or knowing the value of one measurement does not alter or influence the probability of observing any other measurement in the population. After one value is selected, the next value is sampled again at random without regard to the previous measurement, and so on. By contrast, groundwater samples are not chosen at random times or at random locations. The locations are fixed and typically few in number. The intervals between sampling events are fixed and fairly regular. While samples of independent data exhibit no *pairwise correlation* (*i.e.*, no statistical association of similarity or dissimilarity between pairs of sampled measurements), *non-independent* or *dependent* data *do* exhibit pairwise correlation and often other, more complex forms of correlation. Aliquot split sample pairs are generally not independent because of the *positive correlation* induced by the splitting of the same physical groundwater sample. Split measurements tend to be highly similar, much more so than the random pairings of data from distinct sampling events.

In a similar vein, measurements collected close together in time from the same well tend to be more highly correlated than pairs collected at longer intervals. This is especially true when the groundwater is so slow-moving that the same general volume of groundwater is being sampled on closely-spaced consecutive sampling events. Dependence may also be exhibited spatially across a well field. Wells located more closely in space and screened in the same hydrostratigraphic zone may show greater similarity in concentration patterns than wells that are farther apart. For both of these temporal or time-related and spatial dependencies, the observed correlations are a result not only of the non-random nature of the sampling but also the fact that many groundwater populations are not uniform throughout the subsurface. The aquifer may instead exhibit pockets or sub-zones of higher or lower concentration, perhaps due to location-specific differences in natural geochemistry or the dynamics of contaminant plume behavior over time.

As a mathematical construct, statistical independence is essentially impossible to check directly in a set of sample data — other than by ensuring ahead of time that the measurements were collected at random. However, *non-zero pairwise correlation*, a clear sign of dependent data, can be checked and estimated in a variety of ways. The Unified Guidance describes two methods for identifying temporal correlation in **Chapter 14**: the *rank von Neumann ratio* test and the *sample autocorrelation function*. Measurable correlation among consecutive sample pairs may dictate the need for decreasing the sampling frequency or for a more complicated data adjustment.

Defining and modeling wellfield spatial correlation is beyond the scope of this guidance, but is very much the purview of the field of *geostatistics*. The Unified Guidance instead looks for evidence of well-to-well *spatial variation*, *i.e.*, statistically identifiable differences in mean and/or variance levels across the well field. If evident, the statistical approach would need to be modified so that distinct wells are treated as individual populations with statistical testing being conducted separately at each one (*i.e.*, intrawell comparisons).

3.2.2 STATIONARITY

A *stationary* statistical distribution is one whose population characteristics do not change over time and/or space. In a groundwater context, this means that the true population distribution of a given contaminant is the same no matter *where or when* it is sampled. In the strictest form of *stationarity*, the full distribution must be exactly the same at every time and location. However, in practice, a weaker form is usually assumed: that the population mean (μ) and variance (denoted by the Greek symbol σ^2) are the same over time and/or space.

Stationarity is important to groundwater statistical analysis because of the way that monitoring samples must be collected. If a sample set somehow represented the entire population of possible aquifer values, stationarity would not be an issue in theory. A limited number of physical groundwater samples, however, must be individually collected from each sampled location. To generate a statistical sample, the individual measurements must be pooled together over time from multiple sampling events within a well, or pooled together across space by aggregating data from multiple wells, or both.

As long as the contaminant distribution is stationary, such pooling poses no statistical problem. But with a non-stationary distribution, either the mean and/or variance is changing over time in any given well, or the means and variances differ at distinct locations. In either case, the pooled measurements are *not identically-distributed* even if they may be statistically independent.

The effects of non-stationarity are commonly seen in four basic ways in the groundwater context: 1) as spatial variability, 2) in the existence of trends and/or seasonal variation, 3) via other forms of temporal variation, and 4) in the lack of homogeneity of variance. *Spatial variability* (discussed more extensively in **Chapter 13**) refers to statistically identifiable differences in mean and/or variance levels (but usually means) across the well field (*i.e.*, spatial non-stationarity). The existence of such variation often precludes the pooling of data across multiple background wells or the proper upgradient-to-downgradient comparison of background wells against distinct compliance wells. Instead, the usual approach is to perform intrawell comparisons, where well-specific background data is culled from the early sampling history at each well. Checks for spatial variability are conducted graphically with the aid of side-by-side box plots (**Chapter 9**) and through the use of analysis of variance [ANOVA, **Chapter 13**].

A *trend over time* at a given well location indicates that the mean level is not stationary but is instead rising or falling. A *seasonal trend* is similar in that there are periodic increases and decreases. Pooling several sampling events together thus mixes measurements with differing statistical characteristics. This can violate the identically-distributed presumption of almost all statistical tests and usually leads to an inflated estimate of the current population variance. Trends or seasonal variations identified in (upgradient) background wells or in intrawell background data from compliance wells can severely impact the accuracy and effectiveness of statistical procedures described in this guidance if data are pooled over time to establish background limits. The approach that should be taken will vary with the circumstance. Sometimes the trend component might need to be estimated and removed from the original data, so that what gets tested are the *data residuals* (*i.e.*, values that result from subtracting the estimated trend from the original data) instead of the raw measurements. In other cases, an alternate statistical approach might be needed such as a test for (positive) trend or construction of a confidence band around an estimated trend. More discussion of these options is presented in **Chapters 6**, **7**, **14**, and **21**.

To identify a linear trend, the Unified Guidance describes simple linear regression and the Mann-Kendall test in **Chapter 17**. For seasonal patterns or a combination of linear and seasonal trend effects, the guidance discusses the seasonal Mann-Kendall test and the use of ANOVA tests to identify seasonal effects. These diagnostic procedures are also presented in **Chapter 14**.

Temporal variations are distinguished in this guidance from trends or seasonal effects by the lack of a regular or identifiable pattern. Often a temporal effect will be observed as a temporary shift in concentration levels that is similar in magnitude and direction at multiple wells. This can occur at some sites, for instance, due to rainfall or recharge events. Because the mean level changes at least temporarily, pooling data over time again violates the assumption of identically-distributed data. In this case, the temporal effect can be identified by looking for parallel traces on a time series plot of multiple wells and then more formally by performing a *one-way ANOVA for temporal effects*. These procedures are described in **Chapter 14**. Once identified, the residuals from the ANOVA can be used for compliance testing, since the common temporal effect has been removed.

Lastly, *homogeneity of variance* is important in ANOVA tests, which simultaneously evaluates multiple groups of data each representing a sample from a distinct statistical population. In the latter test, well means need not be the same; the reason for performing the test in the first place is to find out whether the means do indeed differ. But the procedure assumes that all the group variances are equal or *homogeneous*. Lack of homogeneity or stationarity in the variances causes the test to be much less effective at discovering differences in the well means. In extreme cases, the concentration levels would have to differ by large amounts before the ANOVA would correctly register a statistical difference. Lack of homogeneity of variance can be identified graphically via the use of side-by-side box plots and then more formally with the use of Levene's test. Both these methods are discussed further in **Chapter 11**. Evidence of unequal variances may necessitate the use of a transformation to stabilize the variance prior to running the ANOVA. It might also preclude use of the ANOVA altogether for compliance testing, but require intrawell approaches to be considered instead.

ANOVA is not the only statistical procedure which assumes homogeneity of variance. Prediction limits and control charts require a similar assumption between background and compliance well data. But if only one new sample measurement is collected per well per evaluation period (*e.g.*, semi-annually) it can be difficult to formally test this assumption with the diagnostic methods cited above. As

an alternative, homogeneity of variance can be periodically tested when a sufficient sample size has been collected for each compliance well (see **Chapter 6**).

3.2.3 LACK OF STATISTICAL OUTLIERS

Many authors have noted that *outliers* — extreme, unusual-looking measurements — are a regular occurrence among groundwater data (Helsel and Hirsch, 2002; Gibbons and Coleman, 2001). Sometimes an outlier results from nothing more than a typographical error on a laboratory data sheet or file. In others, the fault is an incorrectly calibrated measuring device or a piece of equipment that was not properly decontaminated. An unusual measurement might also reflect the sampling of a temporary, local 'hot spot' of higher concentration. In each of these situations, outliers in a statistical context represent values that are inconsistent with the distribution of the remaining measurements. Tests for outliers thus attempt to infer whether the suspected outlier could have reasonably been drawn from the same population as the other measurements, based on the sample data observed up to that point. Statistical methods to help identify potential outliers are discussed in **Chapter 12**, including both *Dixon's* and *Rosner's* tests, as well as references to other methods.

The basic problem with including statistical outliers in analyzing groundwater data is that they do not come from the same distribution as the other measurements in the sample and so fail the identically-distributed presumption of most tests. The consequences can be dramatic, as can be seen for instance when considering *non-parametric prediction limits*. In this testing method, one of the largest values observed in the background data such as the maximum, is often the statistic selected as the prediction limit. If a large outlier is present among the background measurements, the prediction limit may be set to this value despite being unrepresentative of the background population. In effect, it arises from another population, *e.g.*, the 'population' of typographical errors. The prediction limit could then be much higher than warranted based on the observed background data and may provide little if any probability that truly contaminated compliance wells will be identified. The test will then have lower than expected *statistical power*.

Overall, it pays to try to identify possible outliers and to either correct the value(s) if possible, or exclude known outliers from subsequent statistical analysis. It is also possible to select a statistical method that is *resistant* to the presence of outliers, so that the test results are still likely to be accurate even if one or more outliers is unidentified. Examples of this last strategy include setting non-parametric prediction limits to values other than the background maximum using repeat testing (see **Chapter 18**) or using Sen's slope procedure to estimate the rate of change in a linear trend (**Chapter 17**).

3.2.4 NORMALITY

Probability distributions introduced in **Section 3.1** are mathematical models used to approximate or represent the statistical characteristics of populations. Knowing the exact form and defining equation of a probability distribution allows one to assess how likely or unlikely it will be to observe particular measurement values (or ranges of values) when selecting or drawing independent, identically distributed [*i.i.d.*] samples from the associated population. This can be done as follows. In the case of a *continuous distributional model*, a curve can be drawn to represent the probability distribution by plotting probability values along the *y*-axis and measurement or concentration values along the *x*-axis. Since the continuum of *x*-values along this curve is infinite, the probability of occurrence of any single possible value is negligible (*i.e.*, zero), and does not equal the height of the curve. Instead, positive probabilities can be computed for *ranges* of possible values by *summing the area under the distributional curve*

associated with the desired range. Since by definition the total area under any probability distribution curve sums to unity, all probabilities are then numbers between 0 and 1.

Probability distributions form the basic building blocks of all statistical testing procedures. Every test relies on comparing one or more statistics computed from the sample data against a *reference distribution*. The reference distribution is in turn a probability distribution summarizing the expected mathematical behavior of the statistic(s) of interest. A formal statistical test utilizes this reference distribution to make inferences about the sample statistic in terms of two contrasting conditions or hypotheses.

In any event, probability distributions used in statistical testing make differing assumptions about how the underlying population of measurements is distributed. A case in point is simultaneous *prediction limits using retesting* (**Chapter 19**). The first and most common version of this test (Davis and McNichols, 1987) is based on an assumption that the sample data are drawn from a *normal probability distribution*. The normal distribution is the well-known bell-shaped curve, perhaps the single most important and frequently-used distribution in statistical analysis. However, it is not the only one. Bhaumik and Gibbons (2006) proposed similar prediction limits for data drawn from a *gamma distribution* and Cameron (2008) did the same for *Weibull-distributed* measurements. This more recent research demonstrates that prediction limits with similar statistical decision error rates can vary greatly in magnitude, depending on the type of data distribution assumed.

Because many tests make an explicit assumption concerning the distribution represented by the sample data, the form and exact type of distribution often has to be checked using a *goodness-of-fit* test. A goodness-of-fit test assesses how closely the observed sample data resemble a proposed distributional model. Despite the wide variety of probability distributions identified in the statistical literature, only a very few goodness-of-fit tests generally are needed in practice. This is because most tests are based on an assumption of *normally-distributed* or *normal* data. Even when an underlying distribution is not normal, it is often possible to use a mathematical transformation of the raw measurements (*e.g.*, taking the *natural logarithm* or *log* of each value) to *normalize* the data set. The original values can be transformed into a set of numbers that behaves as if drawn from a normal distribution. The transformed values can then be utilized in and analyzed with a *normal-theory test* (*i.e.*, a procedure that assumes the input data are normal).

Specific goodness-of-fit tests for checking and identifying data distributions are found in **Chapter 10** of this guidance. These methods all are designed to check the fit to normality of the sample data. Besides the normal, the *lognormal distribution* is also commonly used as a model for groundwater data. This distribution is not symmetric in shape like the bell-shaped normal curve, nor does it have similar statistical properties. However, a simple *log transformation* of lognormal measurements works to normalize such a data set. The *transformed values* can be tested using one of the standard goodness-of-fit tests of *normality* to confirm that the original data were indeed *lognormal*.

More generally, if a sample shows evidence of *non-normality* using the techniques in **Chapter 10**, the initial remedy is to try and find a suitable *normalizing transformation*. A set of useful possible transformations in this regard has been termed the *ladder of powers* (Helsel and Hirsch, 2002). It includes not only the natural logarithm, but also other mathematical power transformations such as the square root, the cube root, the square, *etc*. If none of these transformations creates an adequately normalized data set, a second approach is to consider what are known as *non-parametric* tests. Normal-

theory and other similar *parametric* statistical procedures assume that the form of the underlying probability distribution is known. They are called parametric because the assumed probability distribution is generally characterized by a small set of *mathematical parameters*. In the case of the normal distribution, the general formula describing its shape and properties is completely specified by two parameters: the *population mean* (μ) and the *population variance* (σ^2). Once values for these quantities are known, the exact distribution representing a particular normal population can be computed or analyzed.

Most parametric tests do not require knowledge of the exact distribution represented by the sample data, but rather just the type of distribution (***e.g.***, normal, lognormal, gamma, Weibull, ***etc***.). In more formal terms, the test assumes knowledge of the *family of distributions* indexed by the characterizing parameters. Every different combination of population mean and variance defines a different normal distribution, yet all belong to the normal family. Nonetheless, there are many data sets for which a known distributional family cannot be identified. Non-parametric methods may then be appropriate, since a known distributional form is not assumed. Non-parametric tests are discussed in various chapters of the Unified Guidance. These tests are typically based on either a *ranking* or an *ordering* of the sample magnitudes in order to assess their statistical performance and accuracy. But even non-parametric tests may make use of a normal approximation to define how expected rankings are distributed.

One other common difficulty in checking for normality among groundwater measurements is the frequent presence of *non-detect* values, known in statistical terms as *left-censored* measurements. The magnitude of these sample concentrations is known only to lie somewhere between zero and the *detection* or *reporting limit*; hence the true concentration is partially 'hidden' or censored on the left-hand side of the numerical concentration scale. Because the most effective normality tests assume that all the sample measurements are known and quantified and not censored, the Unified Guidance suggests two possible approaches in this circumstance. First, it is usually possible to simply assume that the true distributional form of the underlying population cannot be identified, and to instead apply a non-parametric test alternative. This solution is not always ideal, especially when using prediction limits and the background sample size is small, or when using control charts (for which there is no current non-parametric alternative to the Unified Guidance recommended test).

As a second alternative, **Chapter 10** discusses methods for assessing *approximate* normality in the presence of non-detects. If normality can be established, perhaps through a normalizing transformation, **Chapter 15** describes methods for estimating the mean and variance parameters of the specific normal distribution needed for constructing tests (such as prediction limits or control charts), even though the exact value of each non-detect is unknown.

3.3 COMMON STATISTICAL MEASURES

Due to the variety of statistical tests and other methods presented in the Unified Guidance, there are a large number of equations and formulas of relevance to specific situations. The most common statistical measures used in many settings are briefly described below.

Sample mean and standard deviation — the mean of a set of measurements of sample size n is simply the arithmetic average of each of the numbers in the sample (denoted by x_i), described by formula [3.1] below. The sample mean is a common estimate of the center or middle of a statistical distribution. That is, $\bar{x}$ is an estimate of μ, the population mean. The basic formula for the sample standard deviation

is given in equation [3.2]. The sample standard deviation is an estimate of the degree of variability within a distribution, indicating how much the values typically vary from the average value or mean. Thus, the standard deviation *s* is an estimate of the population standard deviation σ. Note that another measure of variability, the sample variance, is simply the square of the standard deviation (denoted by s^2) and serves as an estimate of the population variance σ^2.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad [3.1]$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \qquad [3.2]$$

Coefficient of Variation — for positively-valued measurements, the sample coefficient of variation provides a quick and useful indication of the relative degree of variability within a data set. It is computed as $s/\bar{x}$ and so indicates whether the amount of 'spread' in the sample is small or large relative to the average observed magnitude. Sample coefficients of variation can also be calculated for other distributions such as the logarithmic (see discussion on logarithmic statistics below and **Chapter 10**, **Section 10.4**).

Sample percentile — the *p*th percentile of a sample (denoted as $\tilde{x}_p$) is the value such that $p \times 100\,\%$ of the measurements are no greater than $\tilde{x}_p$, while $(1-p)\times 100\,\%$ of the values are no less than $\tilde{x}_p$. Sample percentiles are computed by making an ordered list of the measurements (termed the *order statistics* of the sample) and either selecting an observed value from the sample that comes closest to satisfying the above definition or interpolating between the pair of sample values closest to the definition if no single value meets it.

Slightly different estimates of the sample percentile are used to perform the interpolation depending on the software package or statistics textbook. The Unified Guidance follows Tukey's (1977) method for computing the lower and upper quartiles (*i.e.*, the 25th and 75th sample percentiles, termed *hinges* by Tukey) when constructing box plots (**Chapter 9**). In that setting, the pair of sample values closest to the desired percentile is simply averaged. Another popular method for more generally computing sample percentiles is to set the rank of the desired order statistic as $k = (n+1) \times p$. If k is not an integer, perform linear interpolation between the pair of ordered sample values with ranks just below and just above k.

Median and interquartile range — the sample median is the 50th percentile of a set of measurements, representing the midpoint of an ordered list of the values. It is usually denoted as $\tilde{x}$ or $\tilde{x}_{.5}$, and represents an alternative estimate of the center of a distribution. The interquartile range [IQR] is the difference between the 75th and 25th sample percentiles, thus equal to $(\tilde{x}_{.75} - \tilde{x}_{.25})$. The IQR offers an alternative estimate of variability in a population, since it represents the measurement range of the middle 50% of the ordered sample values. Both the median and the interquartile range are key statistics used to construct box plots (**Chapter 9**).

The median and interquartile range can be very useful as alternative estimates of data centrality and dispersion to the mean and standard deviation, especially when samples are drawn from a highly skewed (i.e., non-symmetric) distribution or when one or more outliers is present. The table below depicts two data sets, one with an obvious outlier, and demonstrates how these statistical measures compare.

The median and interquartile ranges are not affected by the inclusion of an outlier (perhaps an inadvertent reporting of units in terms of ppb rather than ppm). Large differences between the mean and median, as well as between the standard deviation and interquartile range in the second data set can indicate that an anomalous data point may be present.

Data Set #1	Data Set #2
5	5
10	10
15	15
15	15
15	15
20	20
25	25,000
$\bar{x}$ = 15	$\bar{x}$ > 3,500
$\tilde{x}$ = 15	$\tilde{x}$ = 15
s = 6.5	s > 9,000
IQR = 10	IQR = 10

Log-mean, log-standard deviation and Coefficient of Variation — The lognormal distribution is a frequently-used model in groundwater statistics. When lognormally distributed data are transformed, the normally-distributed measurements can then be input into normal-theory tests. The Unified Guidance frequently makes use of quantities computed on log-transformed values. Two of these quantities, the log-mean and the log-standard deviation, represent the sample mean and standard deviation computed using log-transformed values instead of the raw measurements. Formulas for these quantities — denoted $\bar{y}$ and s_y to distinguish them from the measurement-scale mean ($\bar{x}$) and standard deviation (s) — are given below. Prior to calculating the logarithmic mean and standard deviation, the measurement scale data must first be log-transformed. Taking logarithms of the sample mean ($\bar{x}$) and the sample standard deviation (s) based on the original measurement-scale data, will not give the correct result.

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}\log\left(x_i\right) \qquad [3.3]$$

$$s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\log\left(x_i\right)-\bar{y}\right)^2} \qquad [3.4]$$

A population logarithmic coefficient of variation can be estimated from the logarithmically transformed data as: $CV_{\log} = \sqrt{e^{s_y^2}-1}$. It is based solely on the logarithmic standard deviation, s_y, and represents the intrinsic variability of the untransformed data. .

Sample correlation coefficient — correlation is a common numerical measure of the degree of similarity or linear association between two random variables, say x and y. A variety of statistics are used to estimate the correlation depending on the setting and how much is known about the underlying distributions of x and y. Each measure is typically designed to take on values in the range of –1 to +1, where –1 denotes perfect inverse correlation (*i.e.*, as x increases, y decreases, and vice-versa), while +1 denotes perfect correlation (*i.e.*, x and y increase or decrease together), and 0 denotes no correlation (*i.e.*, x and y behave independently of one another). The most popular measure of linear correlation is Pearson's correlation coefficient (r), which can be computed for a set of n sample pairs (x_i, y_i) as:

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}} \qquad [3.5]$$

3.4 HYPOTHESIS TESTING FRAMEWORK

An important component of statistical analysis involves the testing of competing mathematical models, an activity known as *hypothesis testing*. In hypothesis testing, a formal comparison is made between two mutually exclusive possible statements about reality. Usually these statements concern the type or form of underlying statistical population from which the sample data originated, *i.e.*, either the observed data came from one statistical population or from another, but not both. The sample data are used to judge which statistical model identified by the two hypotheses is most consistent with the collected observations.

Hypothesis testing is similar in nature to what takes place in a criminal trial. Just as one of the two statements in an hypothesis test is judged true and the other false, so the defendant is declared either innocent or guilty. The opposing lawyers each develop their theory or model of the crime and what really happened. The jury must then decide whether the available evidence better supports the prosecution's theory or the defense's explanation. Just as a strong presumption of innocence is given to a criminal defendant, one of the statements in a statistical hypothesis is initially favored over the other. This statement, known as the *null hypothesis* [H_0], is only rejected as false if the sample evidence strongly favors the other side of the hypothesis, known as the *alternative hypothesis* [H_A].

Another important parallel is that the same mistakes which can occur in statistical hypothesis testing are made in criminal trials. In a criminal proceeding, the innocent can falsely be declared guilty or the guilty can wrongly be judged innocent. In the same way, if the null hypothesis [H_0] is a true statement about reality but is rejected in favor of the alternative hypothesis [H_A], a mistake akin to convicting the innocent has occurred. Such a mistake is known in statistical terms as a *false positive* or *Type I error*. If the alternative hypothesis [H_A] is true but is rejected in favor of H_0, the mistake is akin to acquitting the guilty. This mistake is known as a *false negative* or *Type II error*.

In a criminal investigation, the test hypotheses can be reversed. A detective investigating a crime might consider a list of probable suspects as potentially guilty (the null hypothesis [H_0]), until substantial evidence is found to exclude one or more suspects [H_A]. The burden of proof for accepting the alternative hypothesis and the kinds of errors which can result are the opposite from a legal trial.

Certain steps are involved in conducting any statistical hypothesis test. First, the null hypothesis H_0 must be specified and is given presumptive weight in the hypothesis testing framework. The observed sample (or a statistic derived from these data) is assumed to follow a known statistical distribution, consistent with the distributional model used to describe reality under H_0. In groundwater monitoring, a null hypothesis might posit that concentration measurements of benzene, for instance, follow a normal distribution with zero mean. This statement is contrasted against the alternative hypothesis, which is constructed as a competing model of reality. Under H_A, the observed data or statistic follows a different distribution, corresponding to a different distributional model. In the simple example above, H_A might posit that benzene concentrations follow a normal distribution, but this time with a mean no less than 20 ppb, representing a downgradient well that has been contaminated.

Complete descriptions of statistical hypotheses are usually not made. Typically, a shorthand formula is used for the two competing statements. Denoting the true population mean as the Greek letter μ and a possible value of this mean as μ_0, a common specification is:

$$H_0 : \mu \leq \mu_0 \text{ vs. } H_A : \mu > \mu_0 \quad [3.6]$$

This formulation clearly distinguishes between the location (*i.e.*, magnitude) of the population mean μ under the two competing models, but it does not specify the *form* of the underlying population itself. In most parametric tests, as explained in **Section 3.2**, the underlying model is assumed to be the normal distribution, but this is not a necessary condition or the basic assumption in all tests. Note also that a *family* of distributions is specified by the hypothesis, not two individual, specific distributions. Any distribution with a true mean no greater than μ_0 satisfies the null hypothesis, while any distribution from the same family with true mean larger than μ_0 satisfies the alternative hypothesis.

Once the statistical hypothesis has been specified, the next step is to actually collect the data and compute whatever test statistic is required based on the observed measurements and the kind of test. The pattern of the observed measurements or the computed test statistic is then compared with the population model predicted or described under H_0. Because this model is specified as a statistical distribution, it can be used to assign probabilities to different results. If the observed result or pattern occurs with very low probability under the null hypothesis model (*e.g.*, with at most a 5% or 1% chance), one of two outcomes is assumed to have occurred. Either the result is a "chance" fluctuation in the data representing a real but unlikely outcome under H_0, or the null hypothesis was an incorrect model to begin with.

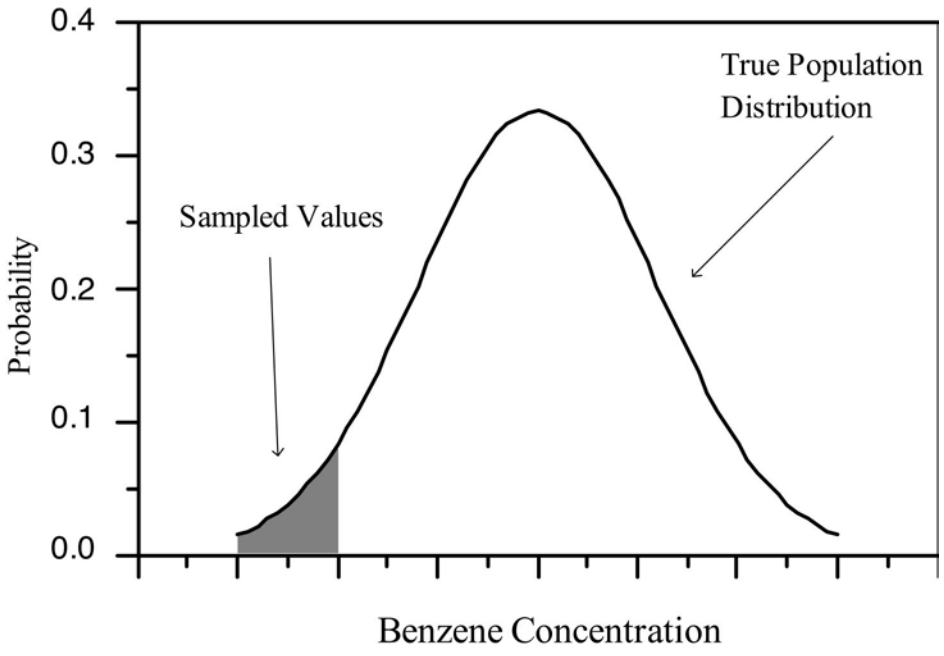
A low probability of occurrence under H_0 is cause for rejecting the null hypothesis in favor of H_A, as long as the probability of occurrence under the latter alternative is also not too small. Still, one should be careful to understand that statistics involves the art of *managing uncertainty*. The null hypothesis may indeed be true, even if the measured results seem unlikely to have arisen under the H_0 model. A *small probability* of occurrence is not the same as *no possibility* of occurrence. The judgment in favor of H_A should be made with full recognition that a *false positive mistake* is always possible even if not very likely.

Consider the measurement of benzene in groundwater in the example above. Given natural fluctuations in groundwater composition from week-to-week or month-to-month and the variability introduced in the lab during the measurement process, the fact that one or two samples show either non-detect or very low levels of benzene does not guarantee that the true mean benzene concentration at the

well is essentially zero. Perhaps the true mean is higher, but the specific sample values collected were gotten from the "lower tail" of the benzene distribution just by chance or were measured incorrectly in the lab. **Figure 3-1** illustrates this possibility, where the full benzene distribution is divided into a lower tail portion that has been sampled and a remaining portion that has not so far been observed. The sampled values are not representative of the entire population distribution, but only of a small part of it.

Along a similar vein, if the observed result or pattern can occur with moderate to high probability under the null hypothesis, the model represented by H_0 is accepted as consistent with the sample measurements. Again, this does not mean the null hypothesis is necessarily true. The alternative hypothesis could be true instead, in which case the judgment to accept H_0 would be considered a *false negative*. Nevertheless the sample data do not provide sufficient evidence or justification to reject the initial presumption.

Figure 3-1. Actual, But Unrepresentative Benzene Measurements



3.5 ERRORS IN HYPOTHESIS TESTING

In order to properly interpret the results of any statistical test, it is important to understand the risks of making a wrong decision. The risks of the two possible errors or mistakes mentioned above are not fixed quantities; rather, false positive and false negative risks are best thought of as statistical parameters that can be adjusted when performing a particular test. This flexibility allows one, in general, to "calibrate" any test to meet specific risk or error criteria. However, it is important to recognize what the different risks represent. RCRA groundwater regulations stipulate that any test procedure maintain a "reasonable balance" between the risks of false positives and false negatives. But how does one decide on a reasonable balance? The answer lies in a proper understanding of the real-life implications attached to wrong judgments.

3.5.1 FALSE POSITIVES AND TYPE I ERRORS

A *false positive* or *Type I error* occurs whenever the null hypothesis [H_0] is falsely rejected in favor of the alternative hypothesis [H_A]. What this means in terms of the underlying statistical models is somewhat different for every test. Many of the tests in the Unified Guidance are designed to address the basic groundwater detection monitoring framework, namely, whether the concentrations at downgradient wells are significantly greater than background. In this case, the null hypothesis is that the background and downgradient wells share the same underlying distribution and that downgradient concentrations should be consistent with background in the absence of any contamination. The alternative hypothesis presumes that downgradient well concentrations are significantly greater than background and come from a distribution with an elevated concentration.

Given this formulation of H_0 and H_A, a Type I error occurs whenever one decides that the groundwater at downgradient locations is significantly higher than background when in reality it is the *same* in distribution. A judgment of this sort concerns the underlying statistical populations and not the observed sample data. The measurements at a downgradient well may indeed be higher than those collected in background. But the disparity must be great enough to decide with confidence that the underlying *populations* also differ. A proper statistical test must account for not just the difference in *observed* mean levels but also variability in the data likely to be present in the underlying statistical populations.

False positive mistakes can cause regulated facilities to incur substantial unnecessary costs and oversight agencies to become unnecessarily involved. Consequently, there is usually a desire by regulators and the regulated community alike to minimize the false positive rate (typically denoted by the Greek letter α). For reasons that will become clear below, the false positive rate is inversely related to the false negative rate for a fixed sample size n. It is impossible to completely eliminate the risk of either Type I or Type II errors, hence the regulatory mandate to minimize the inherent tradeoff by maintaining a "reasonable balance" between false positives and false negatives.

Type I errors are strictly defined in terms of the hypothesis structure of the test. While the conceptual groundwater detection monitoring framework assumes that false positive errors are incorrect judgments of a release when there is none, Type I errors in other statistical tests may have a very different meaning. For instance, in tests of normality (**Chapter 10**) the null hypothesis is that the underlying population is normally-distributed, while the alternative is that the population follows some other, non-normal pattern. In this setting, a false positive represents the mistake of falsely deciding the population to be non-normal, when in fact it *is* normal in distribution. The implication of such an error is quite different, perhaps leading one to select an alternate test method or to needlessly attempt a normalizing transformation of the data.

As a matter of terminology, the false positive rate α is also known as the *significance level* of the test. A test conducted at the $\alpha = .01$ level of significance means there is at most a 1% chance or probability that a Type I error will occur in the results. The test is likely to lead to a false rejection of the null hypothesis at most about 1 out of every 100 times the same test is performed. Note that this last statement says nothing about how well the test will work if H_A is true, when H_0 *should* be rejected. The

false positive rate strictly concerns those cases where H_0 is an accurate reflection of the physical reality, but the test rejects H_0 anyway.

3.5.2 SAMPLING DISTRIBUTIONS, CENTRAL LIMIT THEOREM

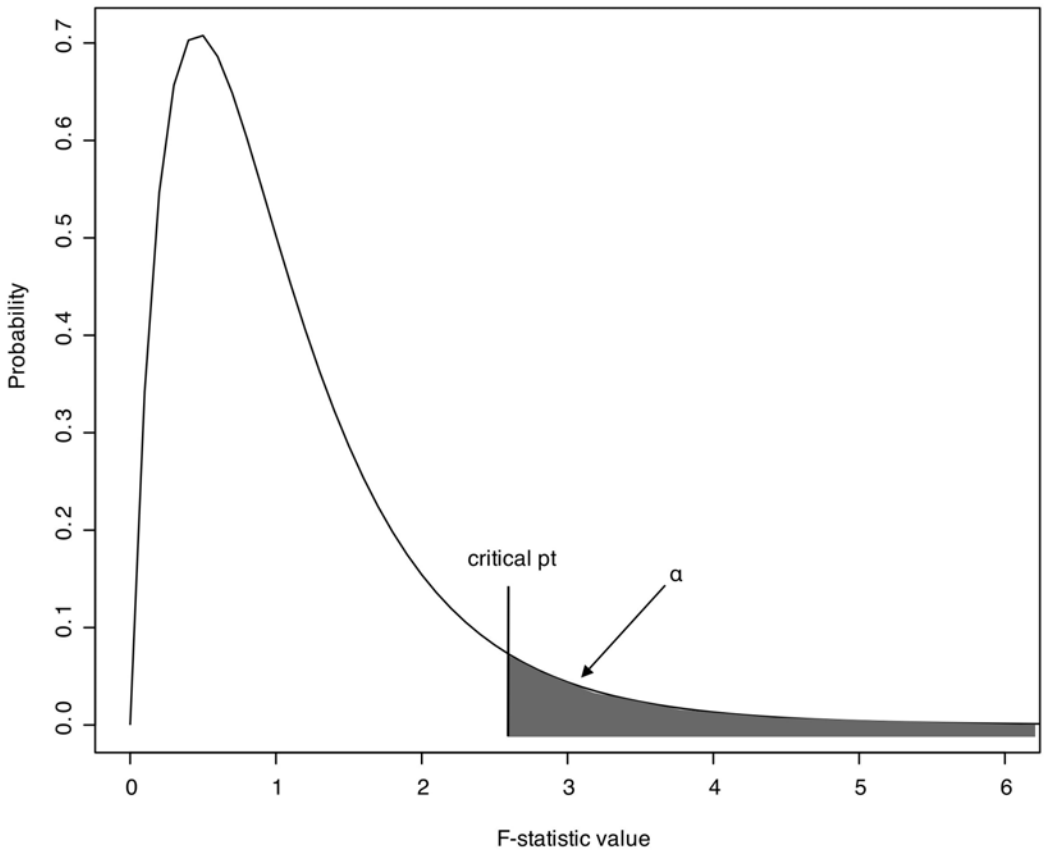
The false positive rate of any statistical test can be calibrated to meet a given risk criterion. To see how this is done, it helps to understand the concept of *sampling distribution*. Most statistical test decisions are based on the magnitude of a particular test statistic computed from the sample data. Sometimes the test statistic is relatively simple, such as the sample mean ($\bar{x}$), while in other instances the statistic is more complex and non-intuitive. In every case, however, the test statistic is formulated as it is for a specific purpose: *to enable the analyst to identify the distributional behavior of the test statistic under the null hypothesis*. Unless one knows the expected behavior of a test statistic, probabilities cannot be assigned to specific outcomes for deciding when the probability is too low to be a chance fluctuation of the data.

The distribution of the test statistic is known as its *sampling distribution*. It is given a special name, in part, to distinguish the behavior of the *test statistic* from the potentially different distribution of the *individual observations or measurements* used to calculate the test. Once identified, the sampling distribution can be used to establish *critical points* of the test associated with specific maximal false positive rates for any given α level of significance. For most tests, a single level of significance is generally chosen.

An example of this idea can be illustrated via the *F*-test. It is used for instance in parametric analysis of variance [ANOVA] to identify differences in the population means at three or more monitoring wells. Although ANOVA assumes that the individual measurements input to the test are normally-distributed, the test statistic under a null hypothesis [H_0] of no differences between the true means follows an *F*-distribution. More specifically, it applies to one member of the *F*-distribution family (an example using 5 wells and 6 measurements per well is pictured in **Figure 3-2**). As seen in the right-hand tail of this distribution by summing the area under the distributional curve, large values of the *F*-statistic become less and less probable as they increase in magnitude. For a given significance level (α), there is a corresponding *F*-statistic value such that the probability of exceeding this cutoff value is α or less. In such situations, there is at most an $\alpha \times 100\%$ chance of observing an F-statistic *under* H_0 that is as large or larger than the cutoff (shaded area in **Figure 3-2**). If α is quite small (*e.g.*, 5% or 1%), one may then judge the null hypothesis to be an untenable model and accept H_A. As a consequence, the cutoff value can be defined as an α-level *critical point* for the *F*-test.

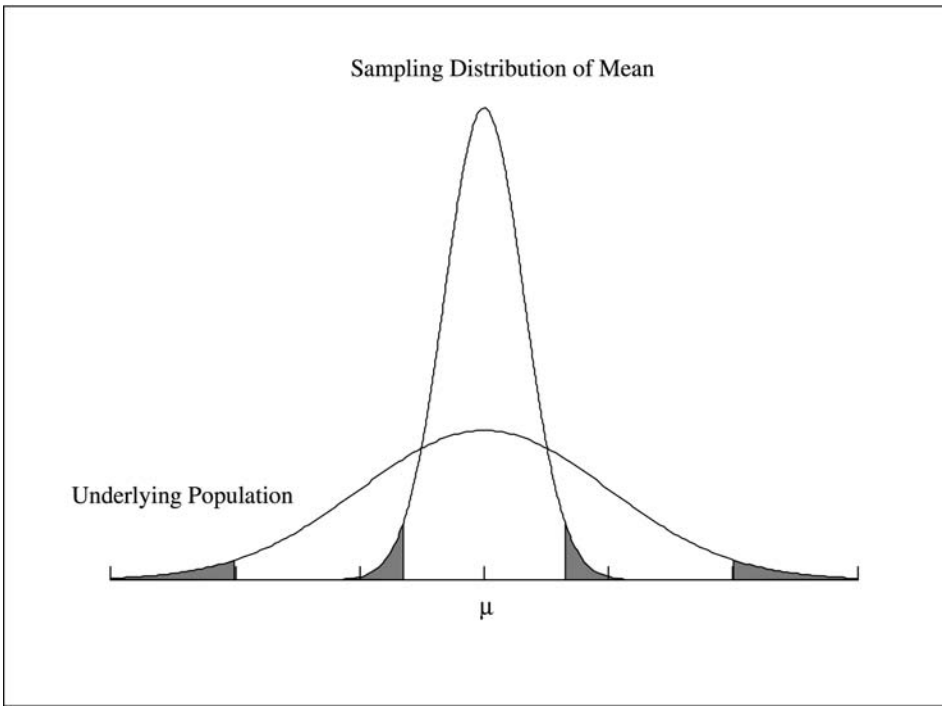
Because test statistics can be quite complicated, there is no easy rule for determining the sampling distribution of a particular test. However, the sampling behavior of some statistics is a consequence of a fundamental result known as the *Central Limit Theorem*. This theorem roughly states that averages or sums of identically-distributed random variables will follow an approximate normal distribution, regardless of the distributional behavior of the individual measurements. This averaged distribution will have the *same* mean μ as the population of individual measurements and whose variance, compared to the underlying population variance σ^2, is scaled by a factor of the sample size *n* on which the average or sum is based. Specifically, the variance is greater by a factor of *n* in the case of a sum $\left(n \cdot \sigma^2\right)$ and smaller by a factor of *n* in the case of an average $\left(\sigma^2/n\right)$. The approximation of the averages or sums to the normal distribution improves as sample size increases (also see the power discussion on page 3-21).

Figure 3-2. F-Distribution with 4 and 25 Degrees of Freedom



Because of the Central Limit Theorem, a number of test statistics at least approximately follow the normal distribution. This allows critical points for these tests to be determined from a table of the standard normal distribution. The Central Limit Theorem also explains why sample means provide a better estimate of the true population mean than individual measurements drawn from the same population (**Figure 3-3**). Since the sampling distribution of the mean is centered on the true average (μ) of the underlying population and the variance is lower by a factor of n, the sample average $\bar{x}$ will tend to be much closer to μ than a typical individual measurement.

Figure 3-3. Effect of Central Limit Theorem



3.5.3 FALSE NEGATIVES, TYPE II ERRORS, AND STATISTICAL POWER

False negatives or Type II errors are the logical opposites of false positive errors. An error of this type occurs whenever the null hypothesis [H_0] is accepted, but instead the alternative hypothesis [H_A] is true. The false negative rate is denoted by the Greek letter β. In terms of the groundwater detection monitoring framework, a Type II error represents a mistake of judging the compliance point concentrations to be consistent with background, when in reality the compliance point distribution is *higher* on average. False negatives in this context describe the risk of *missing* or *not identifying* contaminated groundwater when it really exists. EPA has traditionally been more concerned with such false negative errors, given its mandate to protect human health and the environment.

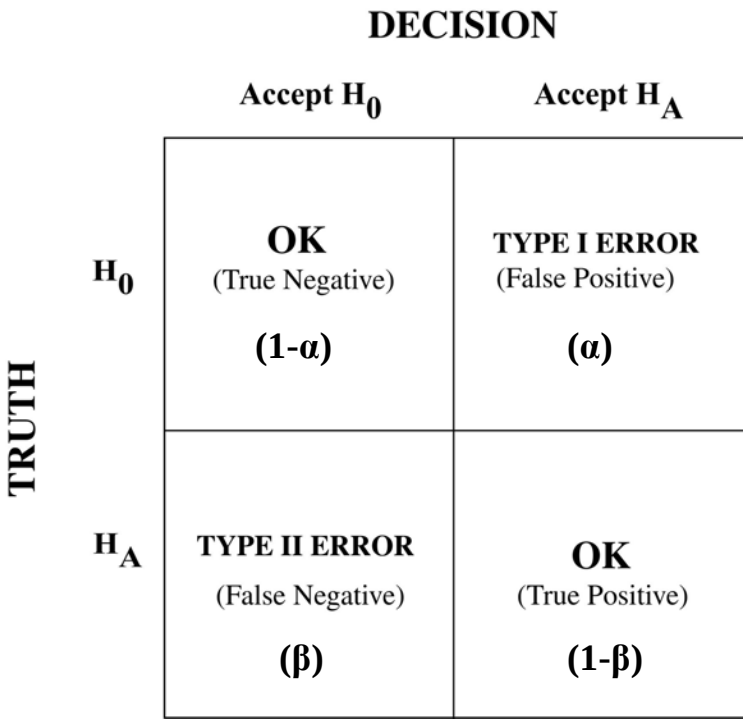
Statistical power is an alternate way of describing false negative errors. Power is merely the complement of the false negative rate. If β is the probability of a false negative, (1–β) is the statistical power of a particular test. In terms of the hypothesis structure, statistical power represents the probability of correctly rejecting the null hypothesis. That is, it is the minimum chance that one will decide to accept H_A, given that H_A is true. High power translates into a greater probability of identifying contaminated groundwater when it really exists.

A convenient way to keep track of the differences between false positives, false negatives, and power is via a Truth Table (**Figure 3-4**). A truth table distinguishes between the *underlying* truth of each hypothesis H_0 or H_A and the *decisions* made on the basis of statistical testing. If H_0 is true, then a decision to accept the alternative hypothesis (H_A) is a false positive error which will occur with a

probability of at most α. Because only one of two decisions is possible, H_0 will also be accepted with a probability of at least (1–α). This is also known as the confidence probability or confidence level of the test, associated with making a 'true negative' *decision*. Similarly if H_A is actually true, making a false negative decision error by accepting the null hypothesis (H_0) has at most a probability of β. Correctly accepting H_A when true then has a probability of at least (1– β) and is labeled a 'true positive' decision. This probability is also known as the statistical power of the test.

For any application of a test to a particular sample, only one of the two types of decision errors can occur. This is because only one of the two mutually exclusive hypotheses will be a true statement. In the detection monitoring context, this means that if a well is *uncontaminated* (*i.e.*, H_0 is true), it may be possible to commit a Type I false positive mistake, but it is *not* possible to make a Type II false negative error. Similarly, if a *contaminated* well is tested (*i.e.*, H_A is true), Type I false positive errors *cannot* occur, but a Type II false negative error might occur.

Figure 3-4. Truth Table in Hypothesis Testing

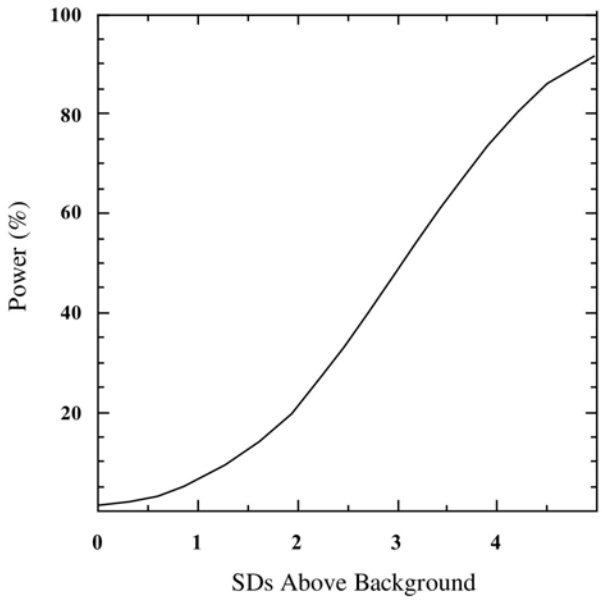


Since the false positive rate can be fixed in advance of running most statistical tests by selecting α, one might think the same could be done with statistical power. Unfortunately, neither statistical power nor the false negative rate can be fixed in advance for a number of reasons. One is that power and the false negative rate depends on the degree to which the true mean concentration level is elevated with respect to the background null condition. Large concentration increases are easier to detect than small increments. In fact, power can be graphed as an increasing function of the true concentration level in what is termed a *power curve* (**Figure 3-5**). A power curve indicates the probability of rejecting H_0 in favor of the alternative H_A for any given alternative to the null hypothesis (*i.e.*, for a range of possible mean-level increases above background).

In interpreting the power curve below, note that the *x*-axis is labeled in terms of relative background standard deviation units (σ) above the true background population mean (μ). The zero point along the *x*-axis is associated with the background mean itself, while the *k*th positive unit along the axis represents a 'true' mean concentration in the compliance well being tested equal to $\mu + k\sigma$. This mode of scaling the graph allows the same power curve to be potentially applied to any constituent of interest subject to the same test conditions. This is true no matter what the typical background concentration levels of a chemical typically found in groundwater may be. But it also means that the same point along the power curve will represent different absolute concentrations for different constituents. Even if the background means are the same, a two standard deviation increase in a chemical with highly variable background concentrations will correspond to a larger population mean increase at a compliance well than the same relative increase in a less variable constituent.

As a simple example, if the background population averages for arsenic and manganese both happen to be 10 ppb, but the arsenic standard deviation is 5 ppb while that for manganese is only 2 ppb, then a compliance well with a mean equivalent to a three standard deviation increase over background would have an average arsenic level of 25 ppb, but an average manganese level of only 16 ppb. For both constituents, however, there would be approximately a 50% probability of detecting a difference between the compliance well and background.

Figure 3-5. Example Power Curve



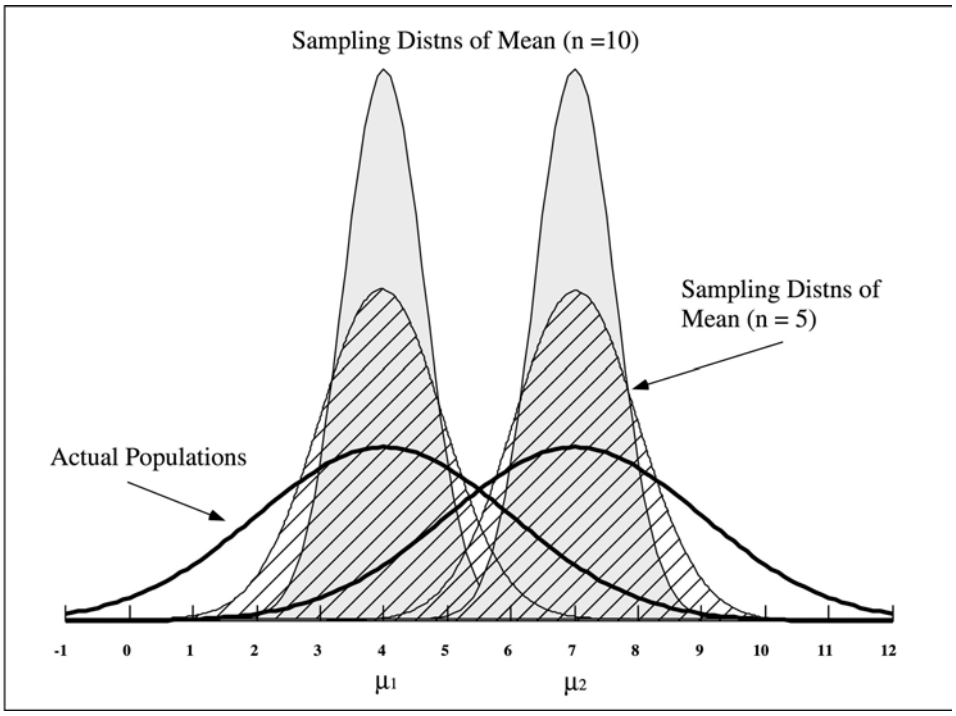
Because the power probability depends on the relative difference between the actual downgradient concentration level and background, power cannot typically be fixed ahead of time like the critical false positive rate for a test. The true concentration level (and associated power) in a compliance well is unknown. If it were known, no hypothesis test would be needed. Additionally, it is often not clear what specific magnitude of increase over background is environmentally significant. A two standard deviation increase over the background average might not be protective of human health and/or the

environment for some monitoring situations. For others, a four standard deviation increase or more may be tolerable before any threat is posed.

Since the exact ramifications of a particular concentration increase are uncertain, it points to the difficulty in setting a minimum power requirement (or a maximum false negative rate) for a given statistical test. Some State statutes contain water quality non-degradation provisions, for which *any* measurable increase might be of concern. By emphasizing relative power as in **Figure 3-5**, all detection monitoring constituents can be evaluated for significant concentration increases on a common footing, subject only to differences in measurement variability.

Another key factor affecting statistical power is sample size. All other test conditions being equal, larger sample sizes provide higher statistical power and the lower the false negative rate (β). Statistical tests perform more accurately with larger data sets, leading to greater power and fewer errors in the process. The Central Limit Theorem illustrates why this is true. Even if a downgradient well mean level is only slightly greater than background, upgradient and downgradient well sample means will have so little variance in their sampling distributions with enough measurements that they will tend to hover very close to their respective population means. True mean differences in the underlying populations can be distinguished with higher probability as sample sizes increase. In **Figure 3-6**, the sampling distributions of means of size 5 and 10 between two different normal populations are provided for illustration. The narrower width of the distribution for the n = 10 sample means are more clearly distinguished from each other than for means of sample size n = 5. This implies higher probability and power to distinguish between the two population means.

Figure 3-6. Why Statistical Power Increases with Sample Size



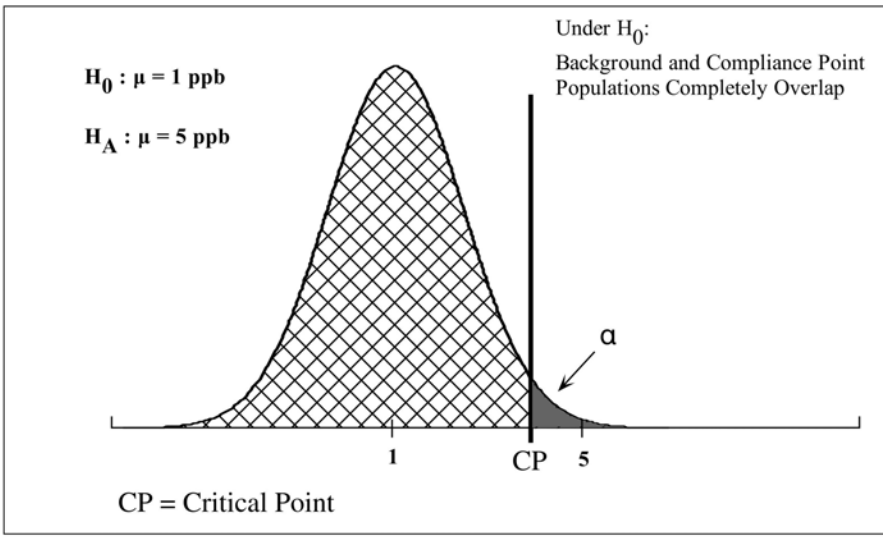
3.5.4 BALANCING TYPE I AND TYPE II ERRORS

In maintaining an appropriate balance between false positive and false negative error rates, one would ideally like to simultaneously minimize both kinds of errors. However, both risks are inherent to any statistical test procedure, and the risk of committing a Type I error is indirectly but inversely related to the risk of a Type II error unless the sample size can be increased. It is necessary to find a *balance* between the two error rates. But given that the false negative rate depends largely on the true compliance point concentrations, it is first necessary to designate what specific mean difference (known as an *effect size*) between the background and compliance point populations should be considered environmentally important. A minimum power requirement can be based on this difference (see **Chapter 6**).

►EXAMPLE 3-1

Consider a simple example of using the downgradient sample mean to test the proposition that the downgradient population mean is 4 ppb larger than background. Assume that extensive sampling has demonstrated that the background population mean is equal to 1 ppb. If the true downgradient mean were the same as the background level, curves of the two sampling distributions would coincide (as depicted in **Figure 3-7**). Then a critical point (e.g., CP = 4.5 ppb) can be selected so that the risk of a false positive mistake is α. The critical point establishes the decision criteria for the test. If the observed sample mean based on randomly selected data from the downgradient sampling distribution exceeds the critical point, the downgradient population will be declared higher in concentration than the background, even though this is not the case. The frequency that such a wrong decision will be made is just the area under the sampling distribution to the right of the critical point equal to α.

Figure 3-7. Relationship Between Type I and Type II Errors, Part A

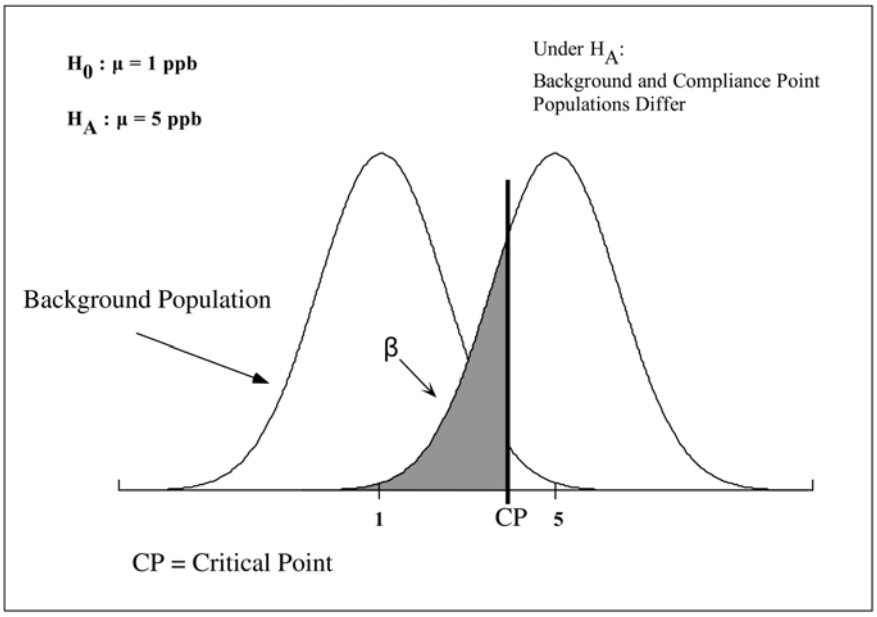


If the true downgradient mean is actually 5 ppb, the sampling distribution of the mean will instead be centered over 5 ppb as in the right-hand curve (*i.e.*, the downgradient population) in **Figure 3-8**. Since there really is a difference between the two populations, the alternative hypothesis and *not* the null

hypothesis is true. Thus, any observed sample mean drawn from the downgradient population then falling below the critical point is a false negative mistake. Consequently, the area under the right-hand sampling distribution in **Figure 3-8** to the *left* of the critical point represents the frequency of Type II errors (β).

The false negative rate (β) in **Figure 3-8** is obviously larger than the false positive rate (α) of **Figure 3-7**. This need not be the case in general, but the key point is to understand that for a fixed sample size, the Type I and Type II error rates cannot be simultaneously minimized. If α is increased, by selecting a lower critical point in **Figure 3-7**, the false negative rate will also be lowered in **Figure 3-8**. Likewise, if α is decreased by selecting a higher critical point, β will be enlarged. If the false positive rate is indiscriminately lowered, the false negative rate (or reduced power) will likely reach unacceptable levels even for mean concentration levels of environmental importance. Such reasoning lay behind EPA's decision to mandate *minimum* false positive rates for *t*-tests and ANOVA procedures in both the revised 1988 and 1991 RCRA rules.

Figure 3-8. Relationship Between Type I and Type II Errors, Part B



◄

This page intentionally left blank

CHAPTER 4. GROUNDWATER MONITORING PROGRAMS AND STATISTICAL ANALYSIS

This chapter provides an overview of the basic groundwater monitoring framework, explaining the intent of the federal groundwater statistical regulations and offering insight into the key identification mechanism of groundwater monitoring, the ***statistically significant increase*** [SSI]:

- What are statistically significant increases and how should they be interpreted?
- What factors, both statistical and non-statistical can cause SSIs?
- What factors should be considered when demonstrating that an SSI does not represent evidence of actual contamination?

4.1 THE GROUNDWATER MONITORING CONTEXT

The RCRA regulations frame a consistent approach to groundwater monitoring, defining the conditions under which statistical testing takes place. Upgradient and downgradient wells must be installed to monitor the uppermost aquifer in order to identify releases or changes in existing conditions as expeditiously as possible. Geological and hydrological expertise is needed to properly locate the monitoring wells in the aquifer passing beneath the monitored unit(s). The regulations identify a variety of design and sampling requirements for groundwater monitoring (such as measuring well piezometric surfaces and identifying flow directions) to assure that this basic goal is achieved. Indicator or hazardous constituents are measured in these wells at regular time intervals; these sample data serve as the basis for statistical comparisons. For identifying releases under detection monitoring, the regulations generally presume comparisons of observations from downgradient wells against those from upgradient wells (designated as background). The rules also recognize certain situations (***e.g.***, mounding effects) when other means to define background may be necessary.

The Unified Guidance may apply to facility groundwater monitoring programs straddling a wide range of conditions. In addition to units regulated under Parts 264 and 265 Subpart F and Part 258 solid waste landfills, other non-regulated units at Subtitle C facilities or CERCLA sites may utilize similar programs. Monitoring can vary from a regulatory minimum of one upgradient and three downgradient wells, to very large facilities with multiple units, and perhaps 50-200 upgradient and downgradient wells. Although the rules presume that monitoring will occur in the single uppermost aquifer likely to be affected by a release, complex geologic conditions may require sampling and evaluating a number of aquifers or strata.

Detection monitoring constituents may include indicators like common ions and other general measures of water quality, pH, specific conductance, total organic carbon [TOC] and total organic halides [TOX]. Quite often, well monitoring data sets are obtained for filtered or unfiltered trace elements (or both) and sizeable suites of hazardous trace organic constituents, including volatiles, semi-volatiles, and pesticide/herbicides. Measurement and analysis of hazardous constituents using standard methods (in SW-846 or elsewhere) have become fairly routine over time. A large number of analytes may be potentially available as monitoring constituents for statistical testing, perhaps 50-100 or more. Identification of the most appropriate constituents for testing depends to a great extent on the composition of the managed wastes (or their decomposition products) as measured in leachate analyses, soil gas sampling, or from prior knowledge.

Nationally, enough groundwater monitoring experience has been gained in using routine constituent lists and analytical techniques to suggest some common underlying patterns. This is particularly true when defining background conditions in groundwater. Sampling frequencies have also been standardized enough (***e.g.***, semi-annual or quarterly sampling) to enable reasonable computation of the sorts of sample sizes that can be used for statistical testing. Nevertheless, complications can and do occur over time — in the form of changes in laboratories, analytical methods, sampled wells, and sampling frequencies — which can affect the quality and availability of sample data.

Facility status can also affect what data are potentially available for evaluation and testing — from lengthy regulated unit monitoring records under the Part 265 interim status requirements at sites awaiting either operational or post-closure 264 permits or permit re-issuance, to a new solid waste facility located in a zone of uncontaminated groundwater with little prior data. Some combined RCRA/CERCLA facilities may have collected groundwater information under differing program requirements. Contamination from offsite or non-regulated units (or solid waste management units) may complicate assessment of likely contaminant sources or contributions.

Quite often, regulators and regulated parties find themselves with considerable amounts of historical constituent-well monitoring data that must be assessed for appropriate action, such as a permit, closure, remedial action or enforcement decision. Users will need to closely consider the diagnostic procedures in **Part II** of the Unified Guidance, with an eye towards selection of one or more appropriate statistical tests in **Parts III** and **IV**. Selection will depend on key factors such as the number of wells and constituents, statistical characteristics of the observed data, and historical patterns of contamination (if present), and may also reflect preferences for certain types of tests. While the Unified Guidance purposely identifies a range of tests which might fit a situation, it is generally recommended that *one* set of tests be selected for final implementation, in order to avoid "test-shopping" (i.e., selecting tests during permit implementation based on the most favorable outcomes). EPA recognizes that the final permit requirements are approved by the regulatory agency.

All of the above situations share some features in common. A certain number of facility wells will be designated as compliance points, ***i.e.***, those locations considered as significant from a regulatory standpoint for assessing potential releases. Similarly, the most appropriate and critical indicator and/or hazardous constituents for monitoring will be identified. If detection monitoring (***i.e.***, comparative evaluations of compliance wells against background) is deemed appropriate for some or all wells and constituents, definitions of background or reference comparison levels will need to be established. Background data can be obtained either from the upgradient wells or from the historical sampling database as described in **Chapter 5**. Choice of background will depend on how statistically comparable

the compliance point data are with respect to background and whether individual constituents exhibit spatial or temporal variability at the facility.

Compliance/assessment or corrective action monitoring may be appropriate choices when there is a prior or historical indication of hazardous constituent releases from a regulated unit. In those situations, the regulatory agency will establish GWPS limits. Typically, these limits are found in established tables, in SDWA drinking water MCLs, through risk-based calculations or determined from background data. For remedial actions, site-specific levels may be developed which account not only for risk, but achievability and implementation costs as well. Nationally, considerable experience has been gathered in identifying cleanup targets which might be applicable at a given facility, as well as how practical those targets are likely to be.

Use of the Unified Guidance should thus be viewed in an overall context. While the guidance offers important considerations and suggestions in selecting and designing a statistically-based approach to monitoring, it is important to realize that it is only a part of the overall decision process at a facility. Geologic and hydrologic expertise, risk-based decisions, and legal and practical considerations by the regulated entity and regulatory agency are fundamental in developing the final design and implementation. The guidance does not attempt to address the many other relevant decisions which impact the full design of a monitoring system.

4.2 RCRA GROUNDWATER MONITORING PROGRAMS

Under the RCRA regulations, some form of statistical testing of sample data will generally be needed to determine whether there has been a release, and if so, whether concentration levels lie below or above a protection standard. The regulations frame the testing programs as detection, compliance/assessment, and corrective action monitoring.

Under RCRA permit development and during routine evaluations, all three monitoring program options may need to be simultaneously considered. Where sufficient hazardous constituent data from site monitoring or other evidence of a release exists, the regulatory agency can evaluate which monitoring program(s) are appropriate under §264.91. Statistical principles and testing provided in the Unified Guidance can be used to develop presumptive evidence for one program over another.

In some applications, more than one monitoring program may be appropriate. Both the number of wells and constituents to be tested can vary among the three monitoring programs at a given site. The types of non-hazardous indicator constituents used for detection monitoring might not be applied in compliance or corrective action monitoring. The latter focus is on hazardous constituents. Only a few compliance well constituents may exceed their respective GWPSs. The focus in a corrective action monitoring program might then be placed on the latter, with the remaining well constituents evaluated under the other monitoring schemes. But following the general regulatory structure, the three monitoring systems are presented below and elsewhere in the guidance as an ordered sequence:

Detection monitoring is appropriate either when there is no evidence of a release from a regulated unit, or when the unit situated in a historically contaminated area is not impacted by current RCRA waste management practices. Care must be taken to avoid a situation where the constituents might reasonably have originated offsite or from units not subject to testing, since any adverse change in groundwater quality would be attributed to on-site causes. Whether an observed change in groundwater

quality is in fact due to a release from on-site waste activities at the facility may be open to dispute and/or further demonstration. However, this basic framework underlies each of the statistical methods used in detection monitoring.

A crucial step in setting up a detection monitoring program is to establish a set of *background* measurements, a baseline or reference level for statistical comparisons (see **Chapter 5**). Groundwater samples from compliance wells are then compared against this baseline to measure changes in groundwater quality. If at least one chemical parameter on the monitoring indicates a *statistically significant increase* above the baseline [SSI, see **Section 4.3**], the facility or regulated unit moves into the next phase: compliance or assessment monitoring.

Compliance or assessment monitoring[1] is appropriate when there is reliable statistical evidence that a concentration increase over the baseline has occurred. The purpose of compliance/assessment monitoring is two-fold: 1) to assess the extent of contamination (*i.e.*, the size of the increase, the chemical parameters involved, and the locations on-site where contamination is evident); and 2) to measure compliance with pre-established numerical concentration limits generally referred to as GWPSs. Only the second purpose is fully addressed using formal statistical tests. While important information can be gleaned from compliance well data, more complex analyses (*e.g.*, contaminant modeling) may be needed to address the first goal.

GWPSs can be fixed health- or risk-based limits, against which single-sample tests are made. At some sites, no specific fixed concentration limit may be assigned or readily available for one or more monitoring parameters. Instead, the comparison is made against a limit developed from background data. In this case, an appropriate statistical approach might be to use the background measurements to compute a statistical limit and set it as the GWPS. See **Chapter 7** for further details. Many of the detection monitoring design principles (**Chapter 6**) and statistical tests (**Part III**) can also be applied to a set of constituents defined by a background-type GWPS.

The RCRA Parts 264 and 258 regulations require an expanded analysis of potential hazardous constituents (Part 258 Appendix II for municipal landfills or Part 264 Appendix IX for hazardous waste units) when detection monitoring indicates a release and compliance monitoring is potentially triggered. The purpose is to better gauge which hazardous constituents have actually impacted groundwater. Some detection monitoring programs may require only limited testing of indicator parameters. This additional sampling can be used to determine which wells have been impacted and provide some understanding of the on-site distribution of hazardous constituent concentrations in groundwater. . The course of action decided by the Regional Administrator or State Director will depend on the number of such chemicals that are present in quantifiable levels and the actual concentration levels.

[1] The terms compliance monitoring (§264.99 & 100) and assessment monitoring (§258.55 & 56) are used interchangeably in this document to refer to RCRA monitoring programs. Compliance monitoring is generally used for permitted hazardous waste facilities under RCRA Subtitle C, while assessment monitoring is applied to municipal solid waste landfills regulated under RCRA Subtitle D. The term "assessment" is also used in 40 CFR 265 Subpart F for a second phase of additional analyte testing. Occasional use is also made of the term "compliance wells," which refers to downgradient monitoring wells located at the point(s) of compliance under §264.95 (any of the three monitoring programs may apply when evaluating these wells).

Following the occurrence of a valid statistically significant increase [SSI] over baseline during detection monitoring, the statistical presumption in compliance/assessment monitoring is quite similar to the detection stage. Given *G* as a fixed compliance or background-derived GWPS, the null hypothesis is that true concentrations (of the underlying compliance point population) are no greater than *G*. This compares to the detection monitoring presumption that concentration levels do not exceed background. One reason for the similarity is that compliance limits may be higher than background levels in some situations. An increase over background in these situations does not necessarily imply an increase over the compliance limit, and the latter must be formally tested. On the other hand, if a health- or risk-based limit is below a background level, the RCRA regulations provide that the GWPS should be based on background.

The Subtitle D regulations for municipal solid waste landfills [MSWLF] stipulate[2] that if "the concentrations of all Appendix II constituents are shown to be at or below background values, using the statistical procedures in §258.53(g), for two consecutive sampling events, the owner or operator... may return to detection monitoring." In other words, assessment monitoring may be exited in favor of detection monitoring when concentrations at the compliance wells are statistically indistinguishable from background for two consecutive sampling periods. While a demonstration that concentration levels are below background would generally not be realistic, it may be possible to show that compliance point levels of contaminants do not exceed an upper limit computed from the background data. Conformance to the limit would then indicate an inability to statistically distinguish between background and compliance point concentration levels.

If a hazardous constituent under compliance or assessment monitoring statistically exceeds a GWPS, the facility is subject to **corrective action.** Remedial activities must be undertaken to remove and/or prevent the further spread of contamination into groundwater. **Monitoring** under corrective action is used to track the progress of remedial activities and to determine if the facility has returned to compliance. Corrective action is usually preceded or accompanied by a formal Remedial Investigation [RI] or RCRA Facility Investigation [RFI] to further delineate the nature and extent of the contaminated plume. Corrective action may be confined to a single regulated unit if only that unit exhibits SSIs above a standard during the detection and compliance/assessment monitoring phases.

Often, clean-up levels are established by the Regional Administrator or State Director during corrective action. Remediation must continue until these clean-up levels are met. The focus of remedial action and monitoring would be on those hazardous constituents and well locations exceeding the GWPSs. If specific clean-up levels have not been met, corrective action must continue until there is evidence of a *statistically significant decrease* [SSD] below the compliance limit for three consecutive years. At this point, corrective action may be exited and compliance monitoring re-started. (As described above and in **Chapter 7**, the protocol for assessing corrective action compliance with a background-type standard can differ). If subsequent concentrations are statistically indistinguishable from background or no detectable concentrations can be demonstrated for three consecutive years in any of the contaminants that triggered corrective measures in the first place, corrective action may be exited in favor of detection monitoring.

[2] [56 FR 51016] October 9, 1991

4.3 STATISTICAL SIGNIFICANCE IN GROUNDWATER TESTING

The outcome of any statistical test is judged either to be statistically significant or non-significant. In groundwater monitoring, a valid statistically significant result can force a change in the monitoring program, perhaps even leading to remedial activity. Consequently, it is important to understand what statistically significant results represent and what they do not. In the language of groundwater hypothesis testing (**Chapter 3**), a statistically significant test result is a decision to reject the null hypothesis (H_0) and to accept the alternative hypothesis (H_A), based on the observed pattern of the sample data. At the most elementary level, a *statistically significant increase* [SSI] (the kind of result typically of interest under RCRA detection and compliance monitoring) represents an observed increase in concentration at one or more compliance wells. In order to be declared an SSI, the change in concentration must be large enough after accounting for variability in the sample data, that the result is unlikely to have occurred merely by chance. What constitutes a statistically significant result depends on the phase of monitoring and the type of statistical test being employed.

If the detection monitoring statistical test being used is a *t*-test or Wilcoxon rank-sum test (**Chapter 16**), an SSI occurs whenever the *t*-statistic or *W*-statistic is larger than an α-level critical point for the test. If a retesting procedure is chosen using a prediction limit (**Chapter 19**), an SSI occurs only when both the initial compliance sample or initial mean/median and one or more resamples all exceed the upper prediction limit. For control charts (**Chapter 20**), an SSI occurs whenever either the CUSUM or Shewhart portions of the chart exceed their respective control limits. In another variation, an SSI only occurs if one or another of the CUSUM or Shewhart statistics exceeds the control limits when recomputed using one or more resamples. For tests of trend (**Chapter 17**), an SSI is declared whenever the slope is significantly greater than zero at some significance level α.

In compliance/assessment monitoring, tests are often made against a fixed compliance limit or GWPS. In this setting, one can utilize a *confidence interval* around a mean, median, upper percentile or a trend line (**Chapter 21**). A confidence interval is an estimated concentration or measurement range intended to contain a given statistical characteristic of the population from which the sample is drawn. A most common formulation is a two-way confidence interval around a normally-distributed mean μ, as shown below:

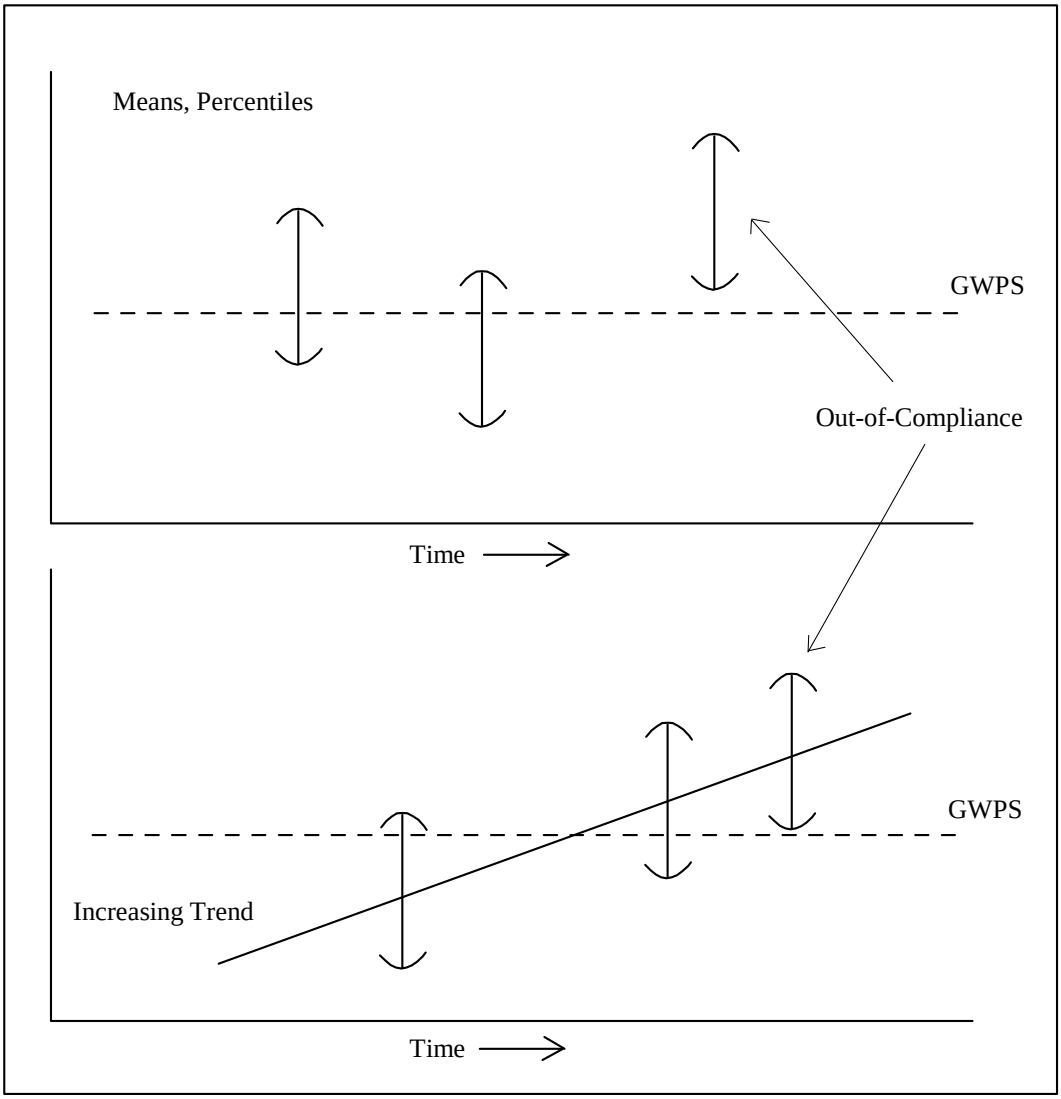
$$\left(\bar{x} - t_{1-\alpha,n-1} \frac{s}{\sqrt{n}} \quad \leq \quad \mu \quad \leq \quad \bar{x} + t_{1-\alpha,n-1} \frac{s}{\sqrt{n}} \right) \qquad [4.1]$$

where $\bar{x}$ is the mean of a sample of size n, s is the sample standard deviation, and $t_{1-\alpha,\, n-1}$ is an upper percentile selected from a Student's *t*-distribution. By constructing a range around the sample mean ($\bar{x}$), this confidence interval is designed to locate the true population mean (μ) with a high degree of *statistical confidence(1-2α)* or conversely, with a low probability of error (2α). If a one-way lower confidence interval is used, the right-hand term in equation [4.1] would be replaced by $+\infty$ at confidence level $1-\alpha$. In a similar fashion, the upper $1-\alpha$ confidence interval would be defined in the range from $-\infty$ for the left-hand term to the right hand term in equation [4.1].

When using a *lower confidence interval* on the mean, median, or upper percentile, an SSI occurs whenever the lower edge of the confidence interval range exceeds the GWPS. For a confidence interval around a trend line, an SSI is declared whenever the lower confidence limit around the estimated trend

line *first* exceeds the GWPS at some point in time. By requiring that a lower confidence limit be used as the basis of comparison, the statistical test will account for data variability and ensure that the apparent violation is unlikely to have occurred by chance. **Figure 4-1** below visually depicts a comparison to a fixed GWPS for both lower confidence intervals for a stationary test like a mean, and around an increasing trend. Where the confidence interval straddles the limit, the test results are inconclusive. In similar fashion, an SSD can be identified by using upper confidence intervals.

Figure 4-1. Confidence Intervals Around Means, Percentiles, or Trend Lines



SSIs offer the primary *statistical* justification for moving from detection monitoring to compliance monitoring, or from compliance/assessment monitoring to corrective action. However, it is important that an SSI be interpreted correctly. Any SSI at a compliance well represents a probable increase in concentration level, but it *does not automatically imply or prove* that contaminated groundwater from the facility is the *cause* of the increase. Due to the complexities of the groundwater medium and the nature of statistical testing, there are numerous reasons why a test may exhibit a statistically significant result. These may or may not be indications of an actual release from a regulated unit.

It is always reasonable to allow for a separate demonstration once an SSI occurs, to determine whether or not the increase is actually due to a contaminant release. Such a demonstration will rely heavily on hydrological and geochemical evidence from the site, but could include additional statistical factors. Key questions and factors to consider are listed in the following sections.

4.3.1 STATISTICAL FACTORS

- Is the result a false positive? That is, were the data tested simply an unusual sample of the underlying population triggering an SSI? Generally, this can be evaluated with repeat sampling.
- Did the test correctly identify an actual release of an indicator or hazardous constituent?
- Are there corresponding SSIs in upgradient or background wells? If so, there may be evidence of a natural in-situ concentration increase, or perhaps migration from an off-site source.
- Is there evidence of significant concentration differences between separate upgradient or background wells, particularly for inorganic constituents? If so, there may be natural spatial variations between distinct well locations that have not been accounted for. These spatial differences could be local or systematic (*e.g.*, upgradient wells in one formation or zone; downgradient wells in another).
- Could observed SSIs for naturally occurring analytes be due to longer-term (*i.e.*, seasonal or multi-year) variation? Seasonal or other cyclical patterns should be observable in upgradient wells. Is this change occurring in both upgradient and downgradient wells? Depending on the statistical test and frequency of sampling involved, an observed SSI may be entirely due to temporal variation not accounted for in the sampling scheme.
- Do time series plots of the sampling data show parallel "spikes" in concentration levels from both background and compliance well samples that were analyzed at about the same time? Perhaps there was an analytical problem or change in lab methodology.
- Are there substantial correlations among within-well constituents (in both upgradient and downgradient wells)? Highly correlated analytes treated as independent monitoring constituents, may generate incorrect significance levels for individual tests.
- Were trends properly accounted for, particularly in the background data?
- Was a correct assumption made concerning the underlying distribution from which the observations were drawn (*e.g.*, was a normal assumption applied to lognormal data)?
- Was the test computed correctly?
- Were the data input to the test of poor quality? (see various factors below)

4.3.2 WELL SYSTEM DESIGN AND SAMPLING FACTORS

- Were early sample data following well installation utilized in statistical testing? Initial well measurements are sometimes highly variable during a 'break in' sampling and analysis period and potentially less trustworthy.
- Was there an effect attributable to recent well development, perhaps due to the use of hazardous constituent chemicals during development or present in drilling muds?
- Are there multiple geological formations at the site, leading to incorrect well placements?

- Has there been degradation of the well casings and screens (*e.g.*, PVC pipe)? Deteriorating PVC materials can release organic constituents under certain conditions. Occasionally, even stainless steel can corrode and release a number of metallic trace elements.
- Have there been changes in well performance over time?
- Were there excessive holding times or incorrect use of preservatives, cooling, *etc.*
- Was there incorrect calibration or drift in the field instrumentation? This effect should be observable in both upgradient and downgradient data and possibly over a number of sample events. The data itself may be compromised or useless.
- Have there been 'mid-stream' changes in sampling procedures, *e.g.*, increased or decreased well purging? Have sampling or purging techniques been consistently applied from well to well or from sampling event to sampling event?

4.3.3 HYDROLOGICAL FACTORS

- Does the site have a history of previous waste management activity (perhaps prior to RCRA), and is there any evidence of historical groundwater contamination? Previous contamination or waste management contaminant levels can limit the ability to distinguish releases from the regulated unit, particularly for those analytes found in historical contamination.
- Is there evidence of groundwater mounding or other anomalies that could lead to the lack of a reliable, definable gradient? Interwell statistical tests assume that changes in downgradient groundwater quality only affect compliance wells and not upgradient (background) wells. Changes that impact background wells also, perhaps in a complex manner involving seasonal fluctuations, are often best resolved by running intrawell tests instead.
- Is there hydrologic evidence of any migration of contaminants (including DNAPL) from off-site sources or from other non-regulated units? Are any of these contaminants observed upgradient of the regulated units?
- Have there been other prior human or site-related waste management activities which could result in the observed SSI changes for certain well locations (*e.g.*, buried waste materials, pipeline leaks, spills, *etc.*)?
- Have there been unusual changes in groundwater directions and depths? Is there confidence that the SSI did indeed correspond to a potential unit release based on observed groundwater directions, distance of the well from the unit, other well information, *etc.*?
- Is there evidence of migration of landfill gas affecting one or more wells?
- Have there been increases in well turbidity and sedimentation, which could affect observed contaminant levels?
- Are there preferential flow paths in the aquifer that could affect where contaminants are likely to be observed or not observed?
- Are the detected contaminants consistent with those found in the waste or leachate of the regulated unit?
- Are there other nearby well pumping or extraction activities?

4.3.4 GEOCHEMICAL FACTORS

- Were the measurements that triggered the SSI developed from unfiltered or filtered trace element sample data? If unfiltered, is there any information regarding associated turbidity or total suspended solid measurements? Unusual increases in well turbidity can introduce excess naturally occurring trace elements into the samples. This can be a particularly difficult problem in compliance monitoring when comparing data to a fixed standard, but can also affect detection monitoring well-to-well comparisons if turbidity levels vary.
- Were there changes in associated analytes at the "triggered" well consistent with local geochemistry? For example, given an SSI for total dissolved solids [TDS], did measured cations/anions and pH also show a consistent change? As another example, slight natural geochemical changes can result in large specific conductance changes. Did other constituents demonstrate a consistent change?
- Is there evidence of a simultaneous release of more than one analyte, consistent with the composition of the waste or leachate? In particular, is there corollary evidence of degradation or daughter products for constituents like halogenated organics? For groundwater constituents with identified SSIs, is there a probable relationship to measured concentrations in waste or waste leachate? Are leachate concentrations high enough to be detectable in groundwater?
- If an SSI is observed in one or more naturally occurring species, were organic hazardous constituents not normally present in background and found in the waste or leachate also detected? This could be an important factor in assessing the source of the possible release.
- Have aquifer mobility factors been considered? Certain soluble constituents like sodium, chloride, or conservative volatile organics might be expected to move through the aquifer much more quickly than easily adsorbed heavy metals or 4-5 ring polynuclear aromatic [PNA] compounds.
- Do the observed data patterns (particularly for naturally occurring constituents in upgradient wells or other background conditions) make sense in an overall site geochemical context, especially as compared with other available local or regional site data and published studies? If not, suspect background data may need to be further evaluated for potential errors prior to formal statistical comparisons.
- Do constituents exhibit correlated behavior among both upgradient and downgradient wells due to overall changes in the aquifer?
- Have there been natural changes in groundwater constituents over time and space due to multi-year, seasonal, or cyclical variation?
- Are there different geochemical regimes in upgradient *vs*. downgradient wells?
- Has there been a release of soil trace elements due to changes in pH?

4.3.5 ANALYTICAL FACTORS

- Have there been changes in laboratories, analytical methods, instrumentation, or procedures including specified detection limits that could cause apparent jumps in concentration levels? In some circumstances, using different values for non-detects with different reporting limits has triggered SSIs. Were inexperienced technicians involved in any of the analyses?

- ❖ Was more than one analytical method used (at different points in time) to generate the measurements?
- ❖ Were there changes in detection/quantification limits for the same constituents?
- ❖ Were there calibration problems, ***e.g.***, drift in instrumentation?
- ❖ Was solvent or other laboratory contamination (***e.g.***, phthalates, methylene chloride extractant, acetone wash) introduced into any of the physical samples?
- ❖ Were there known or probable interferences among the analytes being measured?
- ❖ Were there "spikes" or unusually high values on certain sampling events (either for one constituent among many wells or related analytical constituents) that would suggest laboratory error?

4.3.6 DATA OR ANALYTIC ERRORS

- ❖ Were there data transcription errors (incorrect decimal places, analyte units, or data column entries)? These data can often be identified as being highly improbable.
- ❖ Were there calculation errors in either the analytical (***e.g.***, incorrect trace element valence assumptions or dilution factors) or in the statistical portions (mathematical mistakes, incorrect equation terms) of the analysis?

This page intentionally left blank

CHAPTER 5. ESTABLISHING AND UPDATING BACKGROUND

This chapter discusses the importance and use of background data in groundwater monitoring. Guidance is provided for the proper identification, review, and periodic updating of background. Key questions to be addressed include:

- How should background be established and defined?
- When should existing background data sets be reviewed?
- How and when should background be updated?
- What impact does retesting have on background updating?

5.1 IMPORTANCE OF BACKGROUND

High quality background data is the single most important key to a successful statistical groundwater monitoring program, especially for detection monitoring. All of the statistical tests listed in the RCRA regulations are predicated on having ***appropriate*** and ***representative*** background measurements. As indicated in **Chapter 3**, a statistical sample is representative if the distribution of the sample measurements best follows the distribution of the population from which the sample is drawn. Representative background data has a similar but slightly different connotation. The most important quality of background is that it reflects the historical conditions unaffected by the activities it is designed to be compared to. These conditions could range from an uncontaminated aquifer to an historically contaminated site baseline unaffected by recent RCRA-actionable contaminant releases. Representative background data will therefore have numerical characteristics closely matching those arising from the site-specific aquifer being evaluated.

Background must also be ***appropriate*** to the statistical test. All RCRA detection monitoring tests involve comparisons of compliance point data against background. If natural groundwater conditions

have changed over time — perhaps due to cycles of drought and recharge — background measurements from five or ten years ago may not reflect current uncontaminated conditions. Similarly, recent background data obtained using improved analytical methods may not be comparable to older data. In each case, older background data may have to be discarded in favor of more recent measurements in order to construct an *appropriate* comparison. If intrawell tests are utilized due to strong evidence of spatial variability, traditional upgradient well background data will not provide an appropriate comparison. Even if the upgradient measurements are reflective of uncontaminated groundwater, appropriate background data must be obtained from each compliance point well. The main point is that compliance samples should be tested against data which best can represent background conditions now and those likely to occur in the future.

5.1.1 TRACKING NATURAL GROUNDWATER CONDITIONS

Background measurements, especially from upgradient wells, can provide essential information for other than formal statistical testing. For one, background data can be used to gauge mean levels and develop estimates of variability in naturally occurring groundwater constituents. They can also be used to confirm the presence or absence of anthropogenic or non-naturally occurring constituents in the site aquifer. Ongoing sampling of upgradient background wells provides a means of tracking natural groundwater conditions. Changes that occur in parallel between the compliance point and background wells may signal site-wide aquifer changes in groundwater quality not specifically attributable to onsite waste management. Such observed changes may also be indicative of analytical problems due to common artifacts of laboratory analysis (*e.g.*, re-calibration of lab equipment, errors in batch sample handling, *etc*.), as well as indications of groundwater mounding, changes in groundwater gradients and direction, migration of contaminants from other locations or offsite, etc.

Fixed GWPS like maximum contaminant levels [MCLs] may be contemplated for compliance/assessment monitoring or corrective action. Background data analysis is important if it is suspected that naturally occurring levels of the constituent(s) in question are higher than the standards or if a given hazardous constituent does not have a health- or risk-based standard. In the first case, concentrations in downgradient wells may indeed exceed the standard, but may not be attributable to onsite waste management if natural background levels *also* exceed the standard. The Parts 264 and 258 regulations recognize these possibilities, and allow for GWPS to be based on background levels.

5.2 ESTABLISHING AND REVIEWING BACKGROUND

Establishing appropriate background depends on the statistical approach contemplated (*e.g.*, interwell *vs*. intrawell). This section outlines the major considerations concerning how to select and develop background data including monitoring constituents and sample sizes, statistical assumptions, and the presence of data outliers, spatial variation or trends. Expanding and reviewing background data are also discussed.

5.2.1 SELECTING MONITORING CONSTITUENTS AND ADEQUATE SAMPLE SIZES

Due to the cost of management, mobilization, field labor, and especially laboratory analysis, groundwater monitoring can be an expensive endeavor. The most efficient way to limit costs and still meet environmental performance requirements is to minimize the total number of samples which must be sampled and analyzed. This will require tradeoffs between the number of monitoring constituents

chosen, and the frequency of background versus compliance well testing. The number of compliance wells and annual frequency of testing also affect overall costs, but are generally site-specific considerations. By limiting the number of constituents and ensuring adequate background sample sizes, it is possible to select certain statistical tests which help minimize future compliance (and total) sample requirements.

Selection of an appropriate number of detection monitoring constituents should be dictated by the knowledge of waste or waste leachate composition and the corresponding groundwater concentrations. When historical background data are available, constituent choices may be influenced by their statistical characteristics. A few representative constituents or analytes may serve to accurately assess the potential for a release. These constituents should stem from the regulated wastes, be sufficiently mobile, stable and occur at high enough concentrations to be readily detected in the groundwater. Depending on the waste composition, some non-hazardous organic or inorganic indicator analytes may serve the same purpose. The guidance suggests that between 10-15 formal detection monitoring constituents should be adequate for most site conditions. Other constituents can still be reported but not directly incorporated into formal detection monitoring, especially when large simultaneously analyzed suites like ICP-trace elements, volatile or semi-volatile organics data are run. The focus of adequate background and future compliance test sample sizes can then be limited to the selected monitoring constituents.

The RCRA regulations do not consistently specify how many observations must be collected in background. Under the Part 265 Interim Status regulations, four quarterly background measurements are required during the first year of monitoring. Recent modifications to Part 264 for Subtitle C facilities require a sequence of at least four observations to be collected in background during an interval approved by the Regional Administrator. On the other hand, at least four measurements must be collected from each background well during the first semi-annual period along with at least one additional observation during each subsequent period, for Subtitle D facilities under Part 258. Although these are minimum requirements in the regulations, are they adequate sample sizes for background definition and use?

Four observations from a population are rarely enough to adequately characterize its statistical features; statisticians generally consider sample sizes of $n \leq 4$ to be insufficient for good statistical analysis. A decent population survey, for example, requires several hundred and often a few to several thousand participants to generate accurate results. Clinical trials of medical treatments are usually conducted on dozens to hundreds of patients. In groundwater tests, such large sample sizes are a rare luxury. However, it is feasible to obtain small sample sets of up to $n = 20$ for individual background wells, and potentially larger sample sizes if the data characteristics allow for pooling of multiple well data.

The Unified Guidance recommends that a minimum of at least 8 to 10 independent background observations be collected before running most statistical tests. Although still a small sample size by statistical standards, these levels allow for minimally acceptable estimates of variability and evaluation of trend and goodness-of fit. However, this recommendation should be considered a temporary minimum until additional background sampling can be conducted and the background sample size enlarged (see further discussions below).

Small sample sizes in background can be particularly troublesome, especially in controlling statistical test false positive and negative rates. False negative rates in detection monitoring, *i.e.*, the

statistical error of failing to identify a real concentration increase above background, are in part a function of sample size. For a fixed false positive test rate, a smaller sample size results in a higher false negative rate. This means a decreased probability (*i.e.*, *statistical power*) that real increases above background will be detected. With certain parametric tests, control of the false positive rate using very small sample sets comes at the price of extremely low power. Power may be adequate using a non-parametric test, but control of the false positive can be lost. In both cases, increased background sample sizes result in better achievable false positive and false negative errors.

The overall recommendation of the guidance is to establish background sample sizes as large as feasible. The final tradeoff comes in the selection of the type of detection tests to be used. Prediction limit, control chart, and tolerance limit tests can utilize very small future sample sizes per compliance well (in some cases a single initial sample), but require larger background sample sizes to have sufficient power. Since background samples generally are obtained from historical data sets (plus future increments as needed), total annual sample sizes (and costs) can be somewhat minimized in the future.

5.2.2 BASIC ASSUMPTIONS ABOUT BACKGROUND

Any background sample should satisfy the key statistical assumptions described in **Chapter 3**. These include statistical independence of the background measurements, temporal and spatial stationarity, lack of statistical outliers, and correct distribution assumptions of the background sample when a parametric statistical approach is selected. How independence and autocorrelation impact the establishment of background is presented below, with additional discussions on outliers, spatial variability and trends in the following sections. Stationarity assumptions are considered both in the context of temporal and spatial variation.

Both the Part 264 and 258 groundwater regulations require statistically independent measurements (**Chapter 2**). Statistical *independence* is indicated by random data sets. But randomness is only demonstrated by the presence of mean and variance *stationarity* and the lack of evidence for effects such as *autocorrelation, trends, spatial and temporal variation.* These tests (described in **Part II** of this guidance) generally require at least 8 to 10 separate background measurements.

Depending on site groundwater velocity, too-frequent sampling at any given background well can result in highly *autocorrelated*, non-independent data. Current or proposed sampling frequencies can be tested for autocorrelation or other statistical dependence using the diagnostic procedures in **Chapter 14**. Practically speaking, the best way to ensure some degree of statistical independence is to allow as much time as possible to elapse between sampling events. But a balance must be drawn between collecting as many measurements as possible from a given well over a specified time period, and ensuring that the sample measurements are statistically independent. If significant dependence is identified in already collected background, the interval between sampling events may need to be lengthened to minimize further autocorrelation. With fewer sampling events per evaluation period, it is also possible that a change in statistical method may be needed, say from analysis of variance [ANOVA], which requires at least 4 new background measurements per evaluation, to prediction limits or control charts, which may require new background only periodically (*e.g.*, during a biennial update).

5.2.3 OUTLIERS IN BACKGROUND

Outliers or observations not derived from the same population as the rest of the sample violate the basic statistical assumption of identically-distributed measurements. The Unified Guidance recommends that testing of outliers be performed on background data, but they generally not be removed unless some basis for a likely error or discrepancy can be identified. Such possible errors or discrepancies could include data recording errors, unusual sampling and laboratory procedures or conditions, inconsistent sample turbidity, and values significantly outside the historical ranges of background data. Management of potential outliers carries both positive and negative risks, which should be carefully understood.

If an outlier value with much higher concentration than other background observations is not removed from background prior to statistical testing, it will tend to increase both the background sample mean and standard deviation. In turn, this may substantially raise the magnitude of a parametric prediction limit or control limit calculated from that sample. A subsequent compliance well test against this background limit will be much less likely to identify an exceedance. The same is true with non-parametric prediction limits, especially when the maximum background value is taken as the prediction limit. If the maximum is an outlier not representative of the background population, few truly contaminated compliance wells are likely to be identified by such a test, lowering the statistical power of the method and the overall quality of the statistical monitoring program.

Because of these concerns, it may be advisable at times to remove high-magnitude outliers in background even if the reasons for these apparently extreme observations are not known. The overall impact of removal will tend to improve the power of prediction limits and control charts, and thus result in a more environmentally protective program.

But strategies that involve automated evaluation and removal of outliers may unwittingly eliminate the evidence of real and important changes to background conditions. An example of this phenomenon may have occurred during the 1970s in some early ozone depletion measurements over Antarctica (http://www.nas.nasa.gov/About/Education/Ozone/history.html). Automated computer routines for outlier detection apparently removed several measurements indicating a sharp reduction in ozone concentrations, and thus prevented identification of an enlarging ozone hole by many years. Later review of the raw observations revealed that these automated routines had statistically classified measurements as outliers, which were more extreme than most of the data from that time period. Thus, there is some merit in saving and revisiting apparent 'outliers' in future investigations, even if removed from present databases.

In groundwater data collection and testing, background conditions may not be static over time. Caution should be observed in removing observations which may signal a change in natural groundwater quality. Even when conditions have not changed, an apparently extreme measurement may represent nothing more than a portion of the background distribution that has yet to be observed. This is particularly true if the background data set contains fewer than 20 samples.

In balancing these contrasting risks in retaining or removing one or more outliers, analyses of historical data patterns can sometimes provide more definitive information depending on the types of analytes and methods. For example, if a potential order-of magnitude higher outlier is identified in a sodium data set used as a monitoring constituent, cation-anion balances can help determine if this change is geochemically probable. In this case, changes to other intrawell ions or TDS should be

observed. Similarly, if a trace element outlier is identified in a single well sampling event and occurred simultaneously with other trace element maxima measured using the same analytical method (e.g., ICP-AES) either in the same well or groups of wells, an analytical error should be strongly suspected. On the other hand, an isolated increase without any other evidence could be a real but extreme background measurement. Ideally, removal of one or more statistically identified outliers should be based on other technical information or knowledge which can support that decision.

5.2.4 IMPACT OF SPATIAL VARIABILITY

In the absence of contamination, comparisons made between upgradient-to-downgradient wells assume that the concentration distribution is *spatially stationary* across the well field (**Chapter 3**). This implies that every well should have the same population mean and variance, unless a release occurs to increase the concentration levels at one or more compliance wells. At many sites, this is not the case for many naturally occurring constituents. Natural or man-made differences in mean levels — referred to as *spatial variability* or *spatial variation* — impact how background must be established.

Evidence of spatial variation should drive the selection of an *intrawell* statistical approach if observed among wells known to be uncontaminated (*e.g.*, among a group of upgradient background locations). Lack of spatial mean differences and a common variance allow for *interwell* comparisons. Appropriate background differs between the two approaches.

With interwell tests, background is derived from distinct, initially upgradient background wells, which may be enhanced by data from historical compliance wells also shown not to exhibit significant mean and variance differences. Future data from each of these compliance wells are then tested against this common background. On the other hand, intrawell background is derived from and represents historical groundwater conditions in each individual compliance well. When the population mean levels vary across a well field, there is little likelihood that the upgradient background will provide an appropriate comparison by which to judge any given compliance well.

Although spatial variability impacts the choice of background, it does so *only* for those constituents which evidence spatial differences across the well field. Each monitoring constituent should be evaluated on its own statistical merits. Spatial variation in some constituents (*e.g.*, common ions and inorganic parameters) does not preclude the use of interwell background for other infrequently detected or non-naturally occurring analytes. At many sites, a mixture of statistical approaches may be appropriate: interwell tests for part of the monitoring list and intrawell tests for another portion. Distinct background observation sets will need to be developed under such circumstances.

Intrawell background measurements should be selected from the available historical samples at each compliance well and should include only those observations thought to be uncontaminated. Initially, this might result in very few measurements (*e.g.*, 4 to 6). With such a small background sample, it can be very difficult to develop an adequately powerful intrawell prediction limit or control chart, even when *retesting* is employed (**Chapter 19**). Thus, additional background data will be needed to augment the testing power. One option is to periodically augment the existing background data base with recent compliance well samples (discussed in a further section below). Another possible remedy is to *statistically augment* the available sample data by running an analysis of variance [ANOVA] simultaneously on all the sets of intrawell background from the various upgradient and compliance wells (see **Chapter 13**). The *root mean squared error* [RMSE] from this procedure can be used in place of the

background standard deviation in parametric prediction and control limits to substantially increase the *effective background sample size* of such tests, despite the limited number of observations available per well.

This strategy will only work if the key assumptions of ANOVA can be satisfied (**Chapter 17**), particularly the requirement of equal variances across wells. Since natural differences in mean levels often correspond to similar differences in variability, a transformation of the data will often be necessary to homogenize the variances prior to running the ANOVA. For some constituents, no transformation may work well enough to allow the RMSE to be used as a replacement estimate for the intrawell background standard deviation. In that case, it may not be possible to construct reasonably powerful intrawell background limits until background has been updated once or twice (see **Section 5.3**).

5.2.5 TRENDS IN BACKGROUND

A key implication of the independent and identically distributed assumption [*i.i.d.*] is that a series of sample measurements should be *stationary over time* (*i.e.*, stable in mean level and variance). Data that are trending upward or downward violate this assumption since the mean level is changing. Seasonal fluctuations also violate this assumption since both the mean and variance will likely oscillate. The proper handling of trends in background depends on the statistical approach and the cause of the trend. With interwell tests and a common (upgradient) background, a trend can signify several possibilities:

- Contaminated background;
- A 'break-in' period following new well installation;
- Site-wide changes in the aquifer;
- Seasonal fluctuations, perhaps on the order of several months to a few years.

If upgradient well background becomes contaminated, intrawell testing may be needed to avoid inappropriate comparisons. Groundwater flow patterns should also be re-examined to determine if gradients are properly defined or if groundwater mounding might be occurring. With newly-installed background wells, it may be necessary to discard initially collected observations and to wait several months for aquifer disturbances due to well construction to stabilize. Site-wide changes in the underlying aquifer should be identifiable as similar trends in both upgradient and compliance wells. In this case, it might be possible to remove a common trend from both the background and compliance point wells and to perform interwell testing on the *trend residuals*. However, professional statistical assistance may be needed to do this correctly. Another option would be to switch to intrawell *trend tests* (**Chapter 17**).

Seasonal fluctuations in interwell background which are also observed in compliance wells, can be accommodated by modeling the seasonal trend and removing it from all background and compliance well data. Data seasonally-adjusted in this way (see **Chapter 14** for details) will generally be less variable than the unadjusted measurements and lead to more powerful tests than if the seasonal patterns had been ignored. For this adjustment to work properly, the same seasonal trend should be observed across the well field and not be substantially different from well to well.

Roughly linear trends in *intrawell* background usually signify the need to switch from an intrawell prediction limit or control chart to an explicit trend test, such as *linear regression* or the *Mann-Kendall* (**Chapter 17**). Otherwise the background variance will be overestimated and biased on the high side, leading to higher than expected and ultimately less powerful prediction and control limits. Seasonal fluctuations in intrawell background can be treated in one of two ways. A *seasonal Mann-Kendall* trend test built to accommodate such fluctuations can be employed (**Section 14.3.4**). Otherwise, the seasonal pattern can be estimated and removed from the background data, leaving a set of seasonally-adjusted data to be analyzed with either a prediction limit or control chart. In this latter approach, the same seasonal pattern needs to be extrapolated *beyond* the current background to more recent measurements from the compliance well being tested. These later observations also need to be seasonally-adjusted prior to comparison against the adjusted background, even if there is not enough compliance data yet collected to observe the same seasonal cycles.

When trends are apparent in background, another option is to modify the groundwater monitoring list to include only those constituents that appear to be temporally stable. Only certain analytes may indicate evidence of trends or seasonal fluctuations. More powerful statistical tests might be constructed on constituents that appear to be stationary. All such changes to the monitoring list and method of testing may require approval of the Regional Administrator or State Director.

5.2.6 EXPANDING INITIAL BACKGROUND SAMPLE SIZES

In the initial development of a detection monitoring statistical program under a permit or other legal mechanism, a period of review will identify the appropriate monitoring constituents. For new sites with no prior data, plans for initial background definition need to be developed as part of permit conditions. A more typical situation occurs for interim status or older facilities which have already collected substantial historical data in site monitoring wells. For the most part, the suggestions below cover ways of expanding background data sets from existing information.

Under the RCRA interim status regulations, only a single upgradient well is required as a minimum. Generally speaking, a single background well will not generate observations that are adequately representative of the underlying aquifer. A single background well draws groundwater from only one possible background location. It is accordingly not possible to determine if spatial variation is occurring in the upgradient aquifer. In addition, a single background well can only be sampled so often since measurements that are collected too frequently run the risk of being autocorrelated. Background observations collected from a single well are typically neither representative nor constitute a large enough sample to construct powerful, accurate statistical tests. One way to expand background is to install at least 3-4 upgradient wells and collect additional data under permit.

The early RCRA regulations also allowed for aliquot replicate sampling as a means of expanding background and other well sample sizes. This approach consisted of analyzing splits or aliquots of single water quality samples. As indicated in **Chapter 2**, this approach is not recommended in the guidance. Generally limited analytical variability does not adequately capture the overall variation based on independent water quality sample data, and results in incorrect estimates of variability and degrees of freedom (a function of sample size).

Existing historical groundwater well data under consideration will need to meet the assumptions discussed earlier in this chapter– independence, stationarity, etc., including using statistical methods

which can deal with outliers, spatial and temporal variation including trends. Presuming these conditions are met, it is statistically desirable to develop as large a background sample size as practical. But no matter how many measurements are utilized, a larger sample size is advantageous only if the background samples are both appropriate to the tests selected and representative of baseline conditions.

In limited situations, upgradient-to-downgradient, interwell comparisons may be determined to be appropriate using ANOVA testing of well mean differences. To ensure appropriate and representative background, other conditions may also need to be satisfied when data from separate wells are pooled. First, each background well should be screened at the same hydrostratigraphic position as other background wells. Second, the groundwater chemistry at each of these wells should be similar. This can be checked via the use of standard geochemical bar charts, pie charts, and tri-linear diagrams of the major constituent groundwater ions and cations (Hem, 1989). Third, the *statistical* characteristics of the background wells should be similar — that is, they should be ***spatially stationary***, with approximately the same means and variances. These conditions are particularly important for major water quality indicators, which generally reflect aquifer-specific characteristics. For infrequently detected analytes (***e.g.***, filtered trace elements like chromium, silver, and zinc), even data collected from wells from *different* aquifers and/or geologic strata may be statistically indistinguishable and also eligible for pooling on an interwell basis.

If a one-way ANOVA (**Chapter 13**) on the set of background wells finds significant differences in the mean levels for some constituents, and hence, evidence of spatial variability, the guidance recommends using intrawell tests. The data gathered from the background wells will generally not be used in formal statistical testing, but are still invaluable in ensuring that appropriate background is selected.[1] As indicated in the discussions above and **Chapter 13**, it may be possible to pool constituent data from a number of upgradient and/or compliance wells having a common variance when parametric assumptions allow, even if mean differences exist.

When larger historical databases are available, the data can be reviewed and diagnostically tested to determine which observations best represent natural groundwater conditions suitable for future comparisons. During this review, ***all historical well data*** collected from both upgradient and compliance wells can be evaluated for potential inclusion into background. Wells suspected of prior contamination would need to be excluded, but otherwise each uncontaminated data point adds to the overall statistical picture of background conditions at the site and can be used to enlarge the background database. Measurements can be preferentially selected to establish background samples, so long as a consistent rationale is used (e.g., newer analytical methods, substantial outliers in a portion of a data set, etc.) Changes to an aquifer over time may require selecting newer data representing current groundwater quality over earlier results even if valid.

[1] If the spatial variation is ignored and data are pooled across wells with differing mean levels (and perhaps variances) to run an interwell parametric prediction limit or control chart test, the pooled standard deviation will tend to be substantially larger than expected. This will result in a higher critical limit for the test. Using pooled data with spatial variation will also tends to increase observed maximum values in background, leading to higher and less powerful non-parametric prediction limit tests. In either application, there will be a loss of statistical power for detecting concentration changes at individual compliance wells. Compliance wells with naturally higher mean levels will also be more frequently determined to exceed the limit than expected, while real increases at compliance wells with naturally lower means will go undetected more often.

5.2.7 REVIEW OF BACKGROUND

As mentioned above, if a large historical database is available, a critical review of the data can be undertaken to help establish initially appropriate and representative background samples. We recommend that other reviews of background also take place periodically. These include the following situations:

- When periodically updating background, say every 1-2 years (see **Section 5.3**)
- When performing a 5-10 year permit review

During these reviews, all observations designated as background should be evaluated to ensure that they still adequately reflect current natural or baseline groundwater conditions. In particular, the background samples should be investigated for apparent trends or outliers. Statistical outliers may need to be removed, especially if an error or discrepancy can be identified, so that subsequent compliance tests can be improved. If trends are indicated, a change in the statistical method or approach may be warranted (see earlier section on "Trends in Background").

If background has been updated or enlarged since the last review, and is being utilized in parametric tests, the assumption of normality (or other distributional fit) should be re-checked to ensure that the augmented background data are still consistent with a parametric approach. The presence of non-detects and multiple reporting limits (especially with changes in analytical methods over time) can prove particularly troublesome in checking distribution assumptions. The methods of **Chapters 10** "Fitting Distributions" and **Chapter 15** "Handling Non-Detects" can be consulted for guidance.

Other periodic checks of the revised background should also be conducted, especially in relation to accumulated knowledge from other sites regarding analyte concentration patterns in groundwater. The following are potential sources for comparison and evaluation:

- reliable regional groundwater data studies or investigations from nearby sites;
- published literature; EPA or other agency groundwater databases like STORET;
- knowledge of typical patterns for background inorganic constituents and trace elements. An example is found in **Table 5-1** at the end of this chapter. Typical surface and groundwater levels for filtered trace elements can also be found in the published literature (***e.g.***, Hem, 1989).

Certain common features of routine groundwater monitoring analytes summarized in **Table 5-1** have been observed in Region 8 and other background data sets, which can have implications for statistical applications. Common water quality indicators like cations and anions, pH, TDS, specific conductance are almost always measurable (detectable) and generally have limited within-well variability. These would be more amenable to parametric applications; however, these measurable analytes are also most likely to exhibit well-to-well spatial variation and various kinds of within- and between-well temporal variation including seasonal and annual trends. Many of these within-well analytes are highly correlated, and would not meet the criterion for independent data if simultaneously used as monitoring constituents.

A second level of common indicator analytes– nitrate/nitrite species, fluoride, TOC and TOX– are less frequently detected and subject to more analytical detection instability (higher and lower

detection/quantitation limits). As such, these analyte data are somewhat less reliable. There is less likelihood of temporal variation, although they can exhibit spatial well differences.

Among routinely monitored .45µ-filtered trace elements, different groups stand out. Barium is routinely detected with limited variation within most wells, but does exhibit spatial variation. Arsenic and selenium commonly occur in groundwater as oxyanions, and data can range from virtually non-detectable to always detected in different site wells. The largest group of trace elements can be considered colloidal metals (Sb, Al, Be, Cd, Cr, Co, Fe, Hg, Mn, Pb, Ni, Sn, Tl, V and Zn). While Al, Mn and Fe are more commonly detected, variability is often quite high; well-to-well spatial variability can occur at times. The remaining colloidal metals are solubility-limited in most background groundwater, generally <1 to < 10 µg/l. But even with filtration, some natural colloidal geologic solid materials can often be detected in individual samples. Since naturally occurring Al, Mn and Fe soil solid levels are much higher, the effects on measured groundwater levels are more pronounced and variable. For most of the analytically and solubility-limited colloidal metals, there may not be any discernible well spatial differences. Often these data can be characterized by a site-wide lognormal distribution, and may be possible to pool individual well data to form larger background sizes.

With unfiltered trace element data, it is more difficult to generalize even regarding background data. The method of well sample extraction and the aquifer characteristics will determine how much solids material may be present in the samples. Excessive amounts of sample solids can result in higher levels of detection but also elevated average values and variability even for solubility-limited trace elements. The effect is most clearly seen when TSS is simultaneously collected with unfiltered data. Increases are proportional to the amount of TSS and the natural background levels for trace elements in soil/solid materials. It is recommended that TSS always be simultaneously monitored with unfiltered trace elements.

Most trace organic monitoring constituents are absent or non-detectable under clean background conditions. However, with existing up-gradient sources, it is more difficult to generalize. More soluble constituents like benzene or chlorinated hydrocarbons may be amenable to parametric distributions, but changes in groundwater levels or direction can drastically affect observed levels. For sparingly soluble compounds like polynuclear aromatics (e.g., naphthalene), aquifer effects can result in highly variable data less amenable to statistical applications.

Table 5-1 was based on the use of analytical methods common in the 1990's to the present. Detectable filtered trace element data for the most part were limited by the available analytic techniques, generally SW-846 Method 6010 ICP-AES and select AA (atomic absorption) methods with lower detection limits in the 1-10 ppb range. As newer methods are incorporated (particularly Method 6020 ICP-MS capable of parts-per-trillion detection limits for trace elements), higher quantification frequencies may result in data demonstrating more complex spatial and temporal characteristics. Table 5-1 merely provides a rough guide to where various data patterns might occur. Any extension of these patterns to other facility data sets should be determined by the formal guidance tests in **Part II**.

The background database can also be specially organized and summarized to examine common behavior among related analytes (*e.g.*, filtered trace elements using ICP-AES) either over time or across wells during common sampling events. Parallel time series plots (**Chapter 9**) are very useful in this regard. Groups of related analytes can be graphed on the same set of axes, or groups of nearby wells for the same analyte. With either plot, highly suspect sampling events can be identified if a similar spike in

concentration or other unusual pattern occurs simultaneously at all the wells or in all the analytes. Analytical measurements that appear to be in error might be removed from the background database.

Cation-anion balances and other more sophisticated geochemical analysis programs can also be used to evaluate the reliability of existing water quality background data. A suite of tests like linear or non-parametric correlations, simple or non-parametric ANOVA described in later chapters offer overall methods for evaluating historical data for background suitability.

5.3 UPDATING BACKGROUND

Due both to the complex behavior of groundwater and the need for sufficiently large sample sizes, background once obtained should not be regarded as a single fixed quantity. Background should be sampled regularly throughout the life of the facility, periodically reviewed and revised as necessary. If a site uses traditional, upgradient-to-downgradient comparisons, it might seem that updating of background is conceptually simple: collect new measurements from each background well at each sampling event and add these to the overall background sample. However, significant trends or changes in one or more upgradient wells might indicate problems with individual wells, or be part of a larger site-wide groundwater change. It is worthwhile to consider the following principles for updating, whether interwell or intrawell testing is used.

5.3.1 WHEN TO UPDATE

There are no firm rules on how often to update background data. The Unified Guidance adopts the general principle that updating should occur when enough new measurements have been collected to allow a two-sample statistical comparison between the existing background data and a potential set of newer data. As mentioned in the following section, trend testing might also be used. With quarterly sampling, at least 4 to 8 new measurements should be gathered to enable such a test; this implies that updating would take place every 1-2 years. With semi-annual sampling, the same principle would call for updating every 2-3 years.

Updating should generally not occur more frequently, since adding a new observation to background every one or two sampling rounds does not allow a statistical evaluation of whether the background mean is stationary over time. Enough new data needs to be collected to ensure that a test of means (or medians in the case of non-normal data) can be conducted. Adding individual observations to background can introduce subtle trends that might go undetected and ultimately reduce the statistical power of formal monitoring tests.

Another practical aspect is that when background is updated, all statistical background limits (*e.g.*, prediction and control limits) needs to be recomputed to account for the revised background sample. At complex sites, updating the limits at each well and constituent on the monitoring list may require substantial effort. This includes resetting the cumulative sum [CUSUM] portions of control charts to zero after re-calculating the control limits and prior to additional testing against those limits. Too-frequent updating could thereby reduce the efficacy of control chart tests.

5.3.2 HOW TO UPDATE

Updating background is primarily a concern for intrawell tests, although some of the guidelines apply to interwell data. The common (generally upgradient) interwell background pool can be tested for

trends and/or changes at intervals depending on the sampling frequencies identified above. Those recently collected measurements from the background well(s) can be added to the existing pool if a Student's *t*-test or Wilcoxon rank-sum test (**Chapter 16**) finds no significant difference between the two groups at the $\alpha = 0.01$ level of significance. Individual background wells should also be evaluated in the same manner for their respective newer data. Two-sample tests of the interwell background data are conducted to gauge whether or not background groundwater conditions have changed substantially since the last update, and are ***not*** tests for indicating a potential release under detection monitoring. A significant *t*-test or Wilcoxon rank-sum result should spur a closer investigation and review of the background sample, in order to determine which observations are most representative of the current groundwater conditions.

With intrawell tests using prediction limits or control charts, updating is performed both to enlarge initially small well-specific background samples and to ensure that more recent compliance measurements are not already impacted by a potential release (even if not triggered by the formal detection monitoring tests). A finding of significance using the above two-sample tests means that the most recent data ***should not*** be added to intrawell background. However, the same caveat as above applies: these are not formal tests for determining a potential release and the existing tests and background should continue to be used.

Updating intrawell background should also not occur until at least 4 to 8 new compliance observations have been collected. Further, a potential update is predicated on there being no ***statistically significant increase*** [SSI] recorded for that well constituent, including since the last update. Then a *t*-test or Wilcoxon rank-sum comparison can be conducted at each compliance well between existing intrawell background and the potential set of newer background. A non-significant result implies that the newer compliance data can be re-classified as background measurements and added to the existing intrawell background sample. On the other hand, a determination of significance suggests that the compliance observations should be reviewed to determine whether a gradual trend or other change has occurred that was missed by the intervening prediction limit or control chart tests. If intrawell tests make use of a common pooled variance, the assumption of equal variance in the pooled wells should also be checked with the newer data.

Some users may wish to evaluate historical and future background data for potential trends. If plots of data versus time suggest either an overall trend in the combined data sets or distinct differences in the respective sets, linear or non-parametric trend tests covered in **Chapter 17** might be used. A determination of a significant trend might occur even if the two-sample tests are inconclusive, but individual group sample sizes should be large enough to avoid identifying a significant trend based on too few samples and perhaps randomly occurring. A trend in the newer data may reflect or depart from the historical data conditions. Some form of statistical adjustments may be necessary, but see **Section 5.3.4** below.

5.3.3 IMPACT OF RETESTING

A key question when updating intrawell background is how to handle the results of retesting.[2] If a retest confirms an SSI, background should not be updated. Rather, some regulatory action at the site should be taken. But what if an initial exceedance of a prediction or control limit is *disconfirmed* by retesting? According to the logic of retesting (**Chapter 19**), the well passes the compliance test for that evaluation and monitoring should continue as usual. But what should be done with the initial exceedance when it comes time to update background at the well?

The initial exceedance may be due to a laboratory error or other anomaly that has caused the observation to be an outlier. If so, the error should be documented and not included in the updated background sample. But if the exceedance is not explainable as an outlier or error, it may represent a portion of the background population that has heretofore not been sampled. In that case, the data value could be included in the updated background sample (along with the repeat sample) as evidence of the expanded but true range of background variation. Ultimately, it is important to characterize the background conditions at the site as completely and accurately as possible, so as to minimize both false positive and false negative decision errors in compliance testing.

The severity and classification of the initial exceedance will depend on the specific retesting strategy that has been implemented (**Chapter 19**). Using the same background data in a parametric prediction limit or control chart test, background limits are proportionately lower as the 1-of-*m* order increases (higher *m*). Thus, a 1-of-4 prediction limit will be lower than a 1-of-3 limit, and similarly the 1-of-3 limit lower than for a 1-of-2 test. An initial exceedance triggered by a 1-of-4 test limit and disconfirmed by a repeat sample, might not trigger a lower order prediction limit test. The initial sample value may represent an upper tail value from the true distribution. Retesting schemes derive much of their statistical power by allowing more frequent initial exceedances, even if some of these represent possible measurements from background. The initial and subsequent resamples *taken together* are designed to identify which initial exceedances truly represent SSIs and which do not. These tests presume that occasional excursions beyond the background limit will occur. Unless the exceedance can be documented as an outlier or other anomaly, it should probably be included in the updated intrawell background sample.

5.3.4 UPDATING WHEN TRENDS ARE APPARENT

An increasing or decreasing trend may be apparent between the existing background and the newer set of candidate background values, either using a time series plot or applying **Chapter 17** trend analyses. Should such trend data be added to the existing background sample? Most detection monitoring tests assume that background is stationary over time, with no discernible trends or seasonal variation. A mild trend will probably make very little difference, especially if a Student-*t* or Wilcoxon rank-sum test between the existing and candidate background data sets is non-significant. More severe or continuing trends are likely to be flagged as SSIs by formal intrawell prediction limit or control chart tests.

[2] With interwell tests, the common (upgradient) background is rarely affected by retests at compliance point wells (unless the latter were included in the common pool). Should retesting fail to confirm an initial exceedance , the initial value can be reported alongside the disconfirming resamples in statistical reports for that facility.

With interwell tests, a stronger trend in the common upgradient background may signify a change in natural groundwater quality across the aquifer or an incomplete characterization of the full range of background variation. If a change is evident, it may be necessary to delete some of the earlier background values from the updated background sample, so as to ensure that compliance testing is based on current groundwater conditions and not on outdated measures of groundwater quality.

Table 5-1. Typical Background Data Patterns for Routine Groundwater Monitoring Analytes

Analyte Groups	Detection Rates		Between Well Mean Differences	Within Well Variability (CVs)	Between Well Equal Variances	Outlier Problems	Temporal Variation					Typical Distribution within well	Data Grouping
	Frequency of Detection by Well	Multiple Reporting Limiits					Between Well by Analyte	Within Well Among Group	Within Well Auto-correl.	Within Well Seasonal Variation	Within Well Time Correl.		
Inorganic Constituents and Indicators													
Major ions, pH, TDS, Specific Conductance	High to 100%		✓✓✓	Generally low (.1-.5)	✓✓	✓	✓✓	✓✓✓	✓✓	✓✓	✓✓	Normal	Intrawell
CO3, F, NO2,NO3	Some to most detectable	✓✓	✓✓	Moderate (.2-1.5)	Variable	✓✓	✓			✓	✓	Norm, Log or NPM	Intrawell/ Interwell
.45µ Filtered Trace Elements													
Ba	High to 100%	✓✓	✓✓✓	Low (.1-.5)	✓	✓	✓				✓	Normal	Intrawell
As, Se	Some wells high, others low to zero	✓✓	✓✓ (some wells)	Moderate (.2-1.5)	Variable	✓✓	✓				✓	Normal, Log or NPM	Intrawell/ Interwell
Al, Mn, Fe	Low to Moderate	✓✓	✓	Moderate to high (.3->2.0)	✓	✓✓✓	✓				✓	Log or NPM	Intrawell/ Interwell
Sb, Be, Cd, Cr, Cu, Hg, Pb, Ni, Ag, Tl, V, Zn	Zero to low	✓✓✓		Moderate to high (.5->2.0)	✓✓	✓✓✓	✓	✓✓			✓	Log or NPM	Interwell or NDC
Trace Organic and Indicator Analytes (patterns at sites with prior contamination; generally absent in clean sites)													
VOA's-BETX and Cl-Hydrocarbons	Variable, can be high	✓	Variable by site and wells			✓	Variable by site and specific wells					Normal, Log or NPM	Intrawell, Interwell or NDC
BNAs, Other Trace Organics	Generally low-mod	✓✓	" " " " "			✓	" " " "					" "	" "
Indicators: TOX, TPH, TOC, sulfide	Variable	✓✓	" " " " "			✓✓✓	" " " "					" "	" "

NPM– non-parametric methods; **NDC-** never-detected constituents

Checks: **None–** unknown, absent or infrequently occurring; ✓ – Occasionally; ✓✓ – Frequently; ✓✓✓ – Very Frequently

CHAPTER 6. DETECTION MONITORING PROGRAM DESIGN

6.1 INTRODUCTION

This chapter addresses the *initial statistical design* of a detection monitoring program, prior to routine implementation. It considers what important elements should be specified in site permits, monitoring development plans or during periodic reviews. A good statistical design can be critically important for ensuring that the routine process of detection monitoring meets the broad objective of the RCRA regulations: using statistical testing to accurately evaluate whether or not there is a release to groundwater at one or more compliance wells.

This guidance recommends a *comprehensive* detection monitoring program design, based on two key performance characteristics: adequate *statistical power* and a low predetermined *site-wide false positive rate* [SWFPR]. The design approach presented in **Section 6.2** was developed in response to the *multiple comparisons problem* affecting RCRA and other groundwater detection programs, discussed in **Section 6.2.1**. Greater detail in applying design cumulative false positives and assessing power follows in the next three sub-sections. In **Section 6.3**, consideration is given to data features that impact proper implementation of statistical testing, such as outliers and non-detects, using interwell versus intrawell tests, as well as the presence of spatial variability or trends. **Section 6.4** provides a general discussion of specific detection testing methods listed in the regulations and their appropriate use. Finally, **Section 6.5** applies the design concepts to three hypothetical site examples.

The principles and statistical tests which this chapter covers for a detection monitoring program can also apply to compliance/corrective action monitoring when a background standard is used. Designing a background standards compliance program is discussed in **Chapter 7** (**Section 7.5**).

6.2 ELEMENTS OF THE STATISTICAL PROGRAM DESIGN

6.2.1 THE MULTIPLE COMPARISONS PROBLEM

The foremost goal in detection monitoring is to identify a real release to groundwater when it occurs. Tests must have adequate *statistical power* to identify concentration increases above background. A second critical goal is to avoid *false positive decision errors*, evaluations where one or more wells are falsely declared to be contaminated when in fact their concentration distribution is similar to background. Unfortunately, there is a trade-off (discussed in **Chapter 3**) between maximizing power and minimizing the false positive rate in designing a statistical testing protocol. The statistical power of a given test procedure using a fixed background sample size (*n*) cannot be improved without increasing the risk of false positive error (and vice-versa).

In RCRA and other groundwater detection monitoring programs, most facilities must monitor and test for multiple constituents at all compliance wells one or more times per year. A separate statistical test[1] for each monitoring constituent-compliance well pair is generally conducted semi-annually. Each additional background comparison test increases the accumulative risk of making a false positive mistake, known statistically as the ***multiple comparisons problem***.[2]

The false positive rate α (or *Type I error*) for an individual test is the probability that the test will falsely indicate an exceedance of background. Often, a single fixed low false positive error rate typically found in textbooks or regulation, e.g., $\alpha = .01$ or .05, is applied to each statistical test performed for every well-constituent pair at a facility. Applying such a common false positive rate (α) to each of several tests can result in an acceptable cumulative false positive error if the number of tests is quite small.

But as the number of tests increases, the false positive rate associated with the testing network as a whole (*i.e.*, across all well-constituent pairs) can be surprisingly high. If enough tests are run, at least one test is likely to indicate potential contamination even if a release has not occurred. As an example, if the testing network consists of 20 separate well-constituent pairs and a 99% confidence upper prediction limit is used for each test ($\alpha = .01$), the expected overall *network-wide* false positive rate is about 18%. There is nearly a 1 in 5 chance that one or more tests will *falsely* identify a release to groundwater at uncontaminated wells. For 100 tests and the same statistical procedure, the overall network-wide false positive rate increases to more than 63%, creating additional steps to verify the lack of contamination at falsely triggered wells. This cumulative false positive error is also indicative of <u>at least</u> one well constituent false positive error, but there could be more. Controlling this cumulative false positive error rate is essential in addressing the *multiple comparisons* problem.

---

[1] The number of samples collected may not be the same as the number of statistical tests (e.g., a mean test based on 2 individual samples). It is the number of <u>tests</u> which affect the multiple comparisons problem.

[2] To minimize later confusion, note that the Unified Guidance applies the term "comparison" somewhat differently than most statistical literature. In statistical theory, multiple *tests* are synonymous with multiple *comparisons*, regardless of the kind of statistical test employed. But because of its emphasis on retesting and resampling techniques, the Unified Guidance uses "comparison" in referring to the evaluation of a single sample value or sample statistic against a prediction or control chart limit. In many of the procedures described in **Chapters 19** and **20**, a single statistical test will involve two or more such individual comparisons, yet all the comparisons are part of the same (individual) test.

Three main strategies (or their combination) can be used to counter the excessive cumulative false positive error rate-- 1) the number of tests can be reduced; 2) the individual test false positive rate can be lowered, or 3) the type of statistical test can be changed. A fourth strategy to increase background sample sizes may also be appropriate. Under an initial monitoring design, one usually works with fixed historical sample sizes. However, background data can later be updated in periodic program reviews.

To make use of these strategies, a sufficiently low *target* cumulative SWFPR needs to be initially identified for design purposes. The target cumulative error applies to a certain regular time period. The guidance recommends and uses a value of 10% over a year period of testing. Reasons for this particular choice are discussed in **Section 6.2.2**. These strategies have consequences for the overall test power of a well monitoring network, which are considered following control of the false positive error.

The *number of tests* depends on the number of monitoring constituents, compliance wells and periodic evaluations. Statistical testing on a regular basis can be limited to constituents shown to be *reliable* indicators of a contaminant release (discussed further in **Section 6.2.2)**. Depending on site conditions, some constituents may need to be tested only at wells for a single regulated waste unit, rather than across the entire facility well network. The frequency of evaluation is a program decision, but might be modified in certain circumstances.

Monitoring data for other parameters should still be routinely collected and reported to trace the potential arrival of new chemicals into the groundwater, whether from changes in waste management practices or degradation over time into hazardous daughter products. By limiting *statistically evaluated* constituents to the most useful indicators, the overall number of statistical tests can be reduced to help meet the SWFPR objective. Fewer tests also imply a somewhat higher single test false positive error rate, and therefore an improvement in power.

As a second strategy, the Type I error rate (α_{test}) applied to each individual test can be lowered to meet the SWFPR. Using the *Bonferroni adjustment* (Miller, 1981), the individual test error is designed to limit the overall (or *experiment-wise*) false positive rate α associated with n individual tests by conducting each individual test at an adjusted significance level of $\alpha_{test} = \alpha/n$. Computational details for this approach are provided in a later section.

A full Bonferroni adjustment strategy was neither implemented in previous guidance[3] nor allowed by regulation. However, the principle of partitioning individual test error rates to meet an overall cumulative false positive error target is highly recommended as a design element in this guidance. Because of RCRA regulatory limitations, its application is restricted to certain detection monitoring

[3] A Bonferroni adjustment was recommended in the 1989 **Interim Final Guidance** [IFG] as a *post-hoc* (*i.e.*, 'after the fact') testing strategy for individual background-to-downgradient well comparisons following an analysis of variance [ANOVA]. However, the adjustment does not always effectively limit the risks to the intended 5% false positive error for any ANOVA test. If more than 5 compliance wells are tested, RCRA regulations restrict the single test error rate to a minimum of α = 1% for each of the individual post-hoc tests following the *F*-test. This in effect raises the cumulative ANOVA test risk above 5% and considerably higher with a larger number of tested wells. At least one contaminated well would typically be needed to trigger the initial *F*-test prior to *post-hoc* testing. This fact was also noted in the 1989 IFG. Additionally, RCRA regulations mandate a minimum α error rate of 5% *per constituent* tested with this strategy. For sites with extensive monitoring parameter lists, this means a substantial risk of at least one false positive test result during any statistical evaluation.

tests-- prediction and tolerance limits along with control charts. Where not restricted by regulation, the Bonferroni approach could be used to design workable single-test or post-hoc testing for ANOVAs to meet the overall SWFPR criterion.

Using this strategy of defining individual false positive test rates to meet a cumulative error target, the effect on statistical power is direct. Given a statistical test and fixed sample size, *a lower false positive rate coincides with lower power of the test to detect contamination at the well.* Some improvement in single test power can be gained by increasing background sample sizes at a fixed test error rate. However, the third strategy of utilizing a different or modified statistical test is generally necessary.

This strategy involves choices among certain detection monitoring tests-- prediction limits, control charts and tolerance intervals-- to enhance both power and false positive error control. Except for small sites with a very limited number of tests, any of the three detection monitoring options should incorporate some manner of *retesting*. Through proper design, retesting can simultaneously achieve sufficiently high statistical power while maintaining control of the SWFPR.

RECOMMENDED GUIDANCE CRITERIA

The design of all testing strategies should specifically address the multiple comparisons problem in light of these two fundamental concerns-- an acceptably low false positive site-wide error rate and adequate power. The Unified Guidance accordingly recommends two statistical performance criteria fundamental to good design of a detection monitoring program:

1. **Application of an annual cumulative SWFPR design target, suggested at *10% per year*.**

2. **Use of *EPA reference power curves* [ERPC] to gauge the *cumulative, annual* ability of any individual test to detect contaminated groundwater when it exists. Over the course of a single year assuming normally-distributed background data, any single test performed at the site should have the ability to detect 3 and 4 standard deviation increases above background at specific power levels at least as high as the reference curves**.

False positive rates (or errors) apply both to individual tests and cumulatively to all tests conducted in some time period. Applying the SWFPR annual 10% rate places different sites and state regulatory programs on an equal footing, so that no facility is unfairly burdened by false positive test results. Use of a single overall target allows a proper comparison to be made between alternative test methods in designing a statistical program. Additional details in applying the SWFPR include the following:

- The SWFPR false positive rate should be measured on a *site-wide* basis, partitioned among the total number of annual statistical tests.

- The SWFPR applies to all statistical tests conducted in an *annual* or calendar year period.

- The total number of *annual statistical tests* used in SWFPR calculations depends on the number of valid monitoring constituents, compliance wells and evaluation periods per year. The number of tests may or may not coincide with the number of annual sampling events, for example, if data for a future mean test are collected quarterly and tested semi-annually.

- The Unified Guidance recommends a uniform approach for dealing with monitoring constituents not historically detected in background (e.g., trace organic compounds routinely analyzed in large analytical suites). It is recommended that such constituents *not* be included in SWFPR computations, and an alternate evaluation protocol be used (referred to as the Double Quantification rule) discussed in **Section 6.2.2.**

Statistical power refers to the ability of a test to identify real increases in concentration levels above background (true SSIs). The power of a test is evaluated on population characteristics and represents average behavior defined by repeated or an infinitely large number of samples. Power is reported as a fraction between 0 and 1, representing the probability that the test will identify a *specific level or degree of increase* above background. Statistical power varies with the size of the average population concentration above background-- generally fairly low power to detect small incremental concentrations and substantially increasing power at higher concentrations.

The ERPC describe the cumulative, annual statistical power to detect increasing levels of contamination above a true background mean. These curves are based on specific normal detection monitoring prediction limit tests of single future samples against background conducted once, twice, or four times in a year. Reference curve power is linked to *relative*, not absolute, concentration levels. Actual statistical test power is closely tied to the underlying variability of the concentration measurements. Since individual data set variability will differ by site, constituent, and often by well, the EPA reference power curves provide a generalized ability to estimate power by standardizing variability. By convention, all background concentration data are assumed to follow a standard normal distribution (occasionally referred to in this document as a *Z*-normal distribution) with a true mean $\mu = 0$ and standard deviation $\sigma = 1.0$. Then, increases above background are measured in increasing the k standard deviation units corresponding to $k\sigma$ mean units above baseline. When the background population can be normalized via a transformation, the same normal-based ERPC can be used without loss of generality.

Ideally, actual test power should be assessed using the original concentration data and associated variability, referred to as *effect size* power analysis. The power of any statistical test can be readily computed and compared to the appropriate reference curve, if not analytically, then by Monte Carlo simulation. But the reference power curves laid out in the Unified Guidance offer an important standard by which to judge the adequacy of groundwater statistical programs and tests. They can be universally applied to all RCRA sites and offer a uniform way to assess the environmental and health protection afforded by a particular statistical detection monitoring program.[4]

Consequently, it is recommended that design of any detection monitoring statistical program include an assessment of its ability to meet the power standards set out in the Unified Guidance. The reference power curve approach does not place an undue statistical burden on facility owners or operators, and is believed to be generally protective of human health and the environment.

[4] The ERPCs are specifically intended for comparing background to compliance data in detection monitoring. Power issues in compliance/assessment monitoring and corrective action are considered in **Chapters 7** and **22**.

Principal features of the ERPC approach include the following:

- Reference curves are based on upper 99% prediction limit tests of single future samples against background. The background sample consists of n = 10 measurements, a minimally adequate background sample size typical of RCRA applications. It is assumed that the background sample and compliance well data are normally distributed and from the same population.

- The three reference curves described below are matched to the *annual* frequency of statistical evaluations: one each for quarterly, semi-annual, and annual evaluations. The annual cumulative false positive testing error is maintained at 1%, testing 1, 2, or 4 single future samples annually against the same background. This represents the ability to identify a release to groundwater in at least one of the 1, 2 or 4 tests over the course of a year. Reporting power on an annual basis was chosen to correspond with the application of a cumulative annual SWFPR.

- In the absence of an acceptable *effect size* increase (**Section 6.2.4**), the Unified Guidance recommends that any statistical test provide at least *55-60% annual power to detecting a 3σ* (*i.e.*, 3 standard deviation) increase above the true background mean and at least *80-85% annual power for detecting increases of 4σ*. The percent power criteria change slightly for the respective reference power curves, depending on the annual frequency of statistical evaluations. For normal populations, a 3σ increase above the background average approximately corresponds to the upper 99th percentile of the background distribution, implying better than a 50% chance of detecting such an increase. Likewise, a 4σ increase corresponds to a true mean greater than the upper 99.99th percentile of the background distribution, with better than a 4-in-5 chance of detecting it.

- A single statistical test is not adequately powerful unless its power matches or betters the appropriate reference curve, at least for mean-level increases of 3 to 4 standard deviation units. The same concept can be applied to the overall detection monitoring test design. It is assumed for statistical design purposes that each individual monitoring well and constituent is of equal importance, and assigned a common test false positive error. *Effective power* then measures the overall ability of the statistical program to identify any single constituent release in any well, assuming all remaining constituents and wells are at background levels. If a number of different statistical methods are employed in a single design, effective power can be defined with respect to the *least powerful* of the methods being employed. Applying effective power in this manner would ensure that every well and constituent is evaluated with adequate statistical power to identify potential contamination, not just those where more powerful tests are applied.

- While the Unified Guidance recommends *effective power* as a general approach, other considerations may outweigh statistical thoroughness. Not all wells and constituents are necessarily of equal practical importance. Specific site circumstances may also result in some anomalous weak test power (e.g., a number of missing samples in a background data set for one or more constituents), which might be remedied by eventually increasing background size. The user needs to consider all factors including effective statistical power criteria in assessing the overall strength of a detection monitoring program.

6.2.2 SITE-WIDE FALSE POSITIVE RATES [SWFPR]

In this section, a number of considerations in developing and applying the SWFPR are provided. Following a brief discussion of SWFPR computations, the next section explains the rationale for the 10% design target SWFPR. Additional detail regarding the selection of monitoring constituents follows, and a final discussion of the Double Quantification rule for never-detected constituents is included in the last section.

For cumulative false positive error and SWFPR computations, the following approach is used. A cumulative false positive error rate α_{cum} is calculated as the probability of at least one statistically significant outcome for a total number of tests n_T in a calendar year at a single false positive error rate α_{test} using the properties of the Binomial distribution:

$$\alpha_{cum} = 1 - (1 - \alpha_{test})^{n_T}$$

By rearranging to solve for α_{test}, the 10% design SWFPR (.1) can be substituted for α_{cum} and the needed per-test false positive error rate calculated as:

$$\alpha_{test} = 1 - (.9)^{1/n_T}$$

Although these calculations are relatively straightforward and were used to develop certain κ-factor tables in the Unified Guidance (discussed in **Section 6.5** and in later chapters), a further simplification is possible using the Bonferroni approximation. This assumes that cumulative, annual SWFPR is roughly the additive sum of all the individual test errors. For low false positive rates typical of guidance application, the Bonferroni results are satisfactorily close to the Binomial formula for most design considerations.

Using this principle, the design 10% SWFPR can be partitioned among the potential annual statistical tests at a facility in a number of ways. For facilities with different annual monitoring frequencies, the SWFPR can be divided among quarterly or semi-annual period tests. Given $\alpha_{SWFPR} = .1$ and n_E evaluation periods, the quarterly cumulative false positive target rate α_E at a facility conducting quarterly testing would be $\alpha_E = \alpha_{SWFPR}/n_E = .1/4 = .025$ or 2.5% (and similarly for semi-annual testing). The total or sub-divided SWFPR can likewise be partitioned among dedicated monitoring well groupings at a multi-unit facility or among individual monitoring constituents as needed.

DEVELOPMENT AND RATIONALE FOR THE SWFPR

The existing RCRA Part 264 regulations for parametric or non-parametric analysis of variance [ANOVA] procedures mandate a Type I error of at least 1% for any individual test, and at least 5% overall. Similarly, the RCRA Part 265 regulations require a minimum 1% error for indicator parameter tests. The rationale for minimum false positive requirements is motivated by statistical power. If the Type I error is set too low, the power of the test will be unacceptably low for any given test. EPA was historically not able to specify a minimum level of acceptable power within the RCRA regulations. To do so would require specification of a minimum difference of environmental concern between the null and alternative test hypotheses. Limits on current knowledge about the health and/or environmental effects associated with incremental changes in concentration levels of Part 264 Appendix IX or Part 258 Appendix II constituents greatly complicate this task. Tests of non-hazardous or low-hazard indicators

might have different power requirements than for hazardous constituents. Therefore, minimum false positive rates were adopted for ANOVA-type procedures until more specific guidance could be recommended. EPA's main concern was adequate statistical power to detect real contamination of groundwater, and not enforcing commonly-used false positive test rates.

This emphasis is evident in §264.98(g)(6) and §258.54(c)(3) for detection monitoring and §264.99(i) and §258.55(g)(2) for compliance monitoring. Both pairs of provisions allow the owner or operator to demonstrate that any statistically significant difference between background and compliance point wells or between compliance point wells and the GWPS is an artifact caused by an error in sampling, analysis, statistical evaluation, or natural variation in groundwater chemistry. The rules clearly expect that there will be occasional false positive errors, but existing rules are silent regarding the cumulative frequency of false positives at regulated facilities.

As previously noted, it is essentially impossible to maintain a low cumulative SWFPR for moderate to large monitoring networks if the Type I errors for individual tests must be kept at or above 1%. However, the RCRA regulations do not impose similar false positive error requirements on the remaining control chart, prediction limit and tolerance interval tests. Strategies that incorporate prediction limit or control chart *retesting* can achieve very low individual test false positive rates while maintaining adequate power to detect contamination. Based on prediction limit research in the 1990's and after, it became clear that these alternative methods with suitable retesting could also control the overall cumulative false positive error rate to manageable levels.

This guidance suggests the use of an annual SWFPR of .1 or 10% as a fundamental element of overall detection monitoring design. The choice of a 10% annual SWFPR was made in light of the tradeoffs between false positive control and testing power. An annual period was chosen to put different sized facilities on a common footing regardless of variations in scheduled testing. It is recognized that even with such a limited error rate, the probability of false positive outcomes over a number of years (such as in the lifetime of a 5-10 year permit) will be higher. However, such relatively limited eventualities can be identified and adjusted for, since the RCRA regulations do allow for demonstration of a false positive error. State programs may choose to use a different annual rate such as 5% depending on the circumstances. But **some** predefined SWFPR in a given evaluation period is essential for designing a detection monitoring program, which can then be translated into target individual test rates for any alternative statistical testing strategy.

To implement this recommendation, a given facility should identify its yearly evaluation schedule as quarterly, semi-annual, or annual. This designation is used both to select an appropriate EPA reference power curve by which to gauge acceptable power, and to select prediction limit and control chart multipliers useful in constructing detection monitoring tests. Some of the strategies described in the Unified Guidance in later chapters require that more than one observation per compliance well be collected prior to statistical testing. ***The cumulative, annual false positive rate is linked not to the frequency of sampling but rather to the frequency of statistical evaluation***. When resamples (or verification resamples) are incorporated into a statistical procedure (**Chapter 19**), the individual resample comparisons comprise part of a single test. When a single future mean of ***m*** individual observations is evaluated against a prediction limit, this constitutes a test based on one ***mean comparison.***

NUMBER OF TESTS AND CONSTITUENTS

In designing a detection monitoring program to achieve the target SWFPR, the number of annual statistical tests to be conducted needs to be identified. This number is calculated as the number of distinct monitoring constituents × the number of compliance wells in the network × the number of annual evaluations. Five constituents and 10 well locations statistically evaluated semi-annually constitute 100 annual tests ($5 \times 10 \times 2$), since each distinct well-constituent pair represents a different statistical test that must be evaluated against their respective backgrounds. Even smaller facilities are likely to have a substantial number of such tests, each incrementally adding to the SWFPR.

While the retesting strategies outlined in **Chapters 19** and **20** can aid tremendously in limiting the SWFPR and ensure adequate statistical power, there are practical limits to meeting these goals due to the limited number of groundwater observations that can be collected and/or the number of retests which can feasibly be run. To help balance the risks of false positive and false negative errors, the number of *statistically-tested* monitoring parameters should be limited to constituents thought to be reliable indicators of a contaminant release.

The guidance assumes that data from large suites of trace elements and organics along with a set of inorganic water quality indicators (pH, TDS, common ions, etc.) are routinely collected as part of historical site groundwater monitoring. The number of constituents potentially available for testing can be quite large, perhaps as many as 100 different analytes. At some sites, the full monitoring lists are too large to feasibly limit the SWFPR while maintaining sufficiently high power.

Non-naturally occurring chemicals such as volatile organic compounds [VOC] and semi-volatile organic compounds [SVOC] are often viewed as excellent indicators of groundwater contamination, and are thereby included in the monitoring programs of many facilities. There is a common misperception that the greater the number of VOCs and SVOCs on the monitoring list, the greater the statistical power of the monitoring program. The reasoning is that if none of these chemicals should normally be detected in groundwater — barring a release — testing for more of them ought to improve the chances of identifying contamination.

But including a large suite of VOCs and/or SVOCs among the mix of monitoring parameters ***can be counterproductive*** to the goal of maintaining adequate ***effective power*** for the site as a whole. Because of the trade-off between statistical power and false positive rates (**Chapter 3)**, the power to detect groundwater contamination in one of these wells even with a retesting strategy in place may be fairly low unless background sample sizes are quite large. This is especially true if the regulatory authority only allows for a single retest.

Suppose 40 VOCs and certain inorganic parameters are to be tested semi-annually at 20 compliance wells totaling 1600 annual statistical tests. To maintain a 10% cumulative annual SWFPR, the per-test false positive rate would then need to be set at approximately $\alpha_{test} = .0000625$. If only 10 constituents were selected for formal testing, the per-test rate would be increased to $\alpha_{test} = .00025$. For prediction limits and other detection tests, higher false positive test rates translate to lower κ-factors and improved power.

Some means of reducing the number of tested constituents is generally necessary to design an effective detection monitoring system. Earlier discussions have already suggested one obvious first step,

by eliminating historically non-detected constituents in background from the formal list of detction monitoring constituents (discussed further in the following section). These constituents are still analyzed and informally tested, but do not count against the SWFPR.

Results of waste and leachate testing and possibly soil gas analysis should serve as the initial basis for designating constituents that are reliable leak detection indicators. Such specific constituents actually present in, or derivable from, waste or soil gas samples, should be further evaluated to determine which can be analytically detected a reasonable proportion of the time. This evaluation should include considerations of how soluble and mobile a constituent may be in the underlying aquifer. Additionally, waste or leachate concentrations should be high enough relative to the groundwater levels to allow for adequate detection. By limiting monitoring and statistical tests to fewer parameters with reasonable detection frequencies and that are significant components of the facility's waste, unnecessary statistical tests can be avoided while focusing on the reliable identification of truly contaminated groundwater.

Initial leachate testing should not serve as the sole basis for designating monitoring parameters. At many active hazardous waste facilities and solid waste landfills, the composition of the waste may change over time. Contaminants that initially were all non-detect may not remain so. Because of this possibility, the Unified Guidance recommends that the list of monitoring parameters subject to formal statistical evaluation be periodically reviewed, for example, every three to five years. Additional leachate compositional analysis and testing may be necessary, along with the measurement of constituents not on the monitoring list but of potential health or environmental concern. If previously undetected parameters are discovered in this evaluation, the permit authority should consider revising the monitoring list to reflect those analytes that will best identify potentially contaminated groundwater in the future.

Further reductions are possible in the number of constituents used for formal detection monitoring tests, even among constituents periodically or always detected. EPA's experience at hazardous waste sites and landfills across the country has shown that VOCs and SVOCs detected in a release generally occur in clusters; it is less common to detect only a single constituent at a given location. Statistically, this implies that groups of detected VOCs or SVOCs are likely to be correlated. In effect, the correlated constituents are measuring a release in similar fashion and not providing fully independent measures. At petroleum refinery sites, benzene, toluene, ethylbenzene and xylenes measured in a VOC scan are likely to be detected together Similarly at sites having releases of 1,1,1-trichloroethane, perhaps 10-12 intermediate chlorinated hydrocarbon degradation compounds can form in the aquifer over time. Finally, among water quality indicators like common ions and TDS, there is a great deal of geochemical inter-relatedness. Again, two or three indicators from each of these analyte groups may suffice as detection monitoring constituents.

The overall goal should be to select only the most reliable monitoring constituents for detection monitoring test purposes. Perhaps 10-15 constituents may be a reasonable target, depending on site-specific needs. Those analytes not selected should still continue to be collected and evaluated. In addition to using the informal test to identify previously undetected constituents described in the next section, information on the remaining constituents (e.g., VOCs, SVOCs and trace elements) can still be important in assessing groundwater conditions, including additional confirmation of a detected release.

DOUBLE QUANTIFICATION RULE

From the previous discussion, a full set of site historical monitoring parameters can be split into three distinct groups: a) those reliable indicators and hazardous constituents selected for formal detection monitoring testing and contributing to the SWFPR; b) other analytes which may be occasionally or even frequently detected and will be monitored for general groundwater quality information but not tested; and c) those meeting the "never-detected" criteria. The last group may still be of considerable interest for eventual formal testing, should site or waste management conditions change and new compounds be detected. All background measurements in the "never-detected" group should be non-detects, whether the full historical set or a subgroup considered most representative (e.g., recently collected background measurements using an improved analytical method.[5]). The following rule is suggested to provide a means of evaluating "never-detected" constituents.

The Double Quantification rule implies that statistical tests should be designed for each of the constituents in the first group. Calculations involving the SWFPR should cover these constituents, but *not* include constituents in second and the third '100% non-detect' categories. Any constituent in this third group should be evaluated by the following simple, quasi-statistical rule[6]:

> ***A confirmed exceedance is registered if any well-constituent pair in the '100% non-detect' group exhibits quantified measurements (i.e., at or above the reporting limit [RL]) in two consecutive sample and resample events.***

It is assumed when estimating an SWFPR using the Bonferroni-type adjustment, that each well-constituent test is at *equal risk* for a *specific, definable* false positive error. As a justification for this Double Quantification rule, analytical procedures involved in identifying a reported non-detect value suggest that the error risk is probably much *lower* for most chemicals analyzed as "never-detected." Reporting limits are set high enough so that if a chemical is *not present at all* in the sample, a detected amount will rarely be recorded on the lab sheet. This is particularly the case since method detection limits [MDLs] are often intended as 99% upper prediction limits on the measured signal of an uncontaminated laboratory sample. These limits are then commonly multiplied by a factor of 3 to 10 to determine the RL.

Consequently, a series of measurements for VOCs or SVOCs on samples of uncontaminated groundwater will tend to be listed as a string of non-detects with possibly a very occasional low-level detection. Because the observed measurement levels (*i.e.*, instrument signal levels) are usually known only to the chemist, an approximate prediction limit for the chemical basically has to be set at the RL. However, the true measurement distribution is likely to be clustered much more closely around zero than the RL (**Figure 6-1**), meaning that the false positive rate associated with setting the RL as the prediction

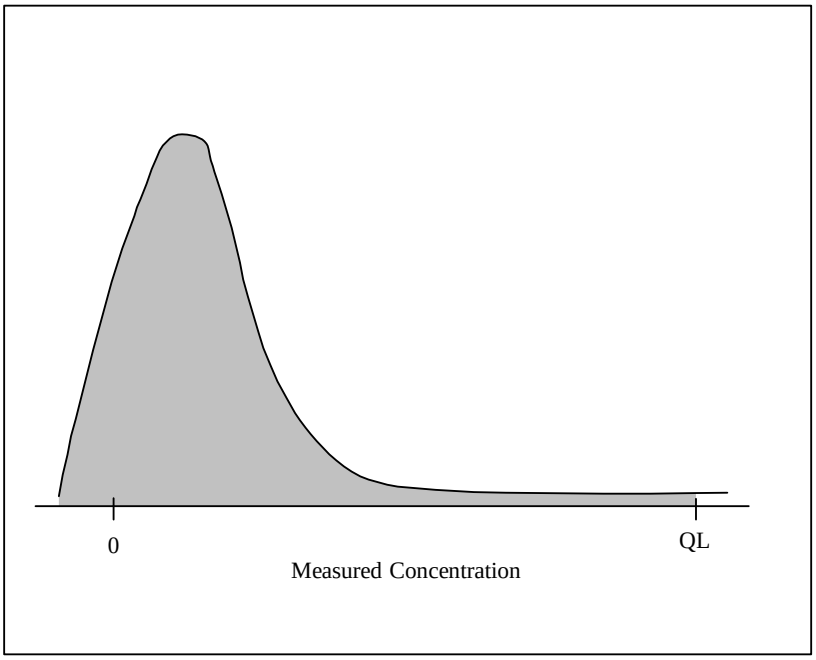
---

[5] Note: Early historical data for some constituents (e.g., certain filtered trace elements) may have indicated occasional and perhaps unusual detected values using older analytical techniques or elevated reporting limits. If more recent sampling exhibits no detections at lower reporting limits for a number of events, the background review discussed in **Chapter 5** may have determined that the newer, more reliable recent data should be used as background. These analytes could also be included in the '100% non-detect' group.

[6] The term "quasi-statistical" indicates that although the form is a statistical prediction limit test, only an approximate false positive error rate is implied for the reporting limit critical value. The test form follows 1-of-2 or 1-of-3 non-parametric prediction limit tests using the maximum value from a background data set (**Chapter 19**).

limit is likely already *much lower* than the Bonferroni-adjusted error rate calculated above. A similar chain of reasoning would apply to site-specific chemicals that may be on the monitoring list but have *never* been detected at the facility. Such constituents would also need a prediction limit set at the RL.

Figure 6-1. Hypothetical Distribution of Instrument Signals in Uncontaminated Groundwater



In general, there should be some minimally sufficient sample numbers to justify placing constituents in the "never-detected" category. Even such a recommendation needs to consider individual background well versus pooled well data. Depending on the number of background wells (including historical compliance well data used as background which reflect the same non-detect patterns), certain risks may have to be taken to implement this strategy. With the same total number of non-detects (e.g., 4 each in 5 wells versus 20 from a single well), the relative risk can change. Certain non-statistical judgements may be needed, such as the likelihood of particular constituents arising from the waste or waste management unit. At a minimum, we recommend that at least 6 consecutive non-detect values initially be present in each well of a pooled group, and additional background well sampling should occur to raise this number to 10-15.

Having 10-15 non-detects as a basis, a maximum worst-case probability of a future false positive exceedance under Double Quantification rule testing could be estimated. But it should be kept in mind that the true individual comparison false positive rates based on analytical considerations are likely to be considerably lower. The number of non-detect constituents evaluated under the rule will also play a role. There will be some cumulative false positive error based on the number of comparisons at some true false positive single test error or errors. Since the true false positive test rates cannot be known (and may vary considerably among analytes), it is somewhat problematic to make this cumulative false positive error estimate. Yet there is some likelihood that occasional false positive exceedances will occur under this rule.

Some flexibility will be required in evaluating such outcomes, particularly if there is doubt that a confirmed exceedance is actually due to a release from the regulated unit. In this circumstance, it might be appropriate to allow for a second resample as more definitive confirmation.

In implementing the Double Quantification rule, consideration should be given to how soon a repeat sample should be taken. Unlike detectable parameters, the question of autocorrelation is immaterial since the compound should not be present in the background aquifer. A sufficiently long interval should occur between the initial and repeat samples to minimize the possibility of a systematic analytical error. But the time interval should be short enough to avoid missing a subsequent real detection due to seasonal changes in the aquifer depth or flow direction. It is suggested that 1-2 months could be appropriate, but will depend on site-specific hydrological conditions.

Using this rule, it should be possible to construct adequately powerful prediction and control limits for naturally-occurring and detectable inorganic and organic chemicals in almost every setting. This is especially helpful at larger sites, since the total number of tests on which the per-test false positive rates (α_{test}) are based will be significantly reduced. Requiring a verified quantification for previously non-detected constituents should ensure that spurious lab results do not falsely trigger a facility into compliance/assessment monitoring, and will more reliably indicate the presence of chemicals that have heretofore not been found in background.

6.2.3 RECOMMENDATIONS FOR STATISTICAL POWER

The second but more important regulatory goal of a testing strategy is to ensure sufficient *statistical power* for detecting contaminated groundwater. Technically, in the context of groundwater monitoring, power refers to the probability that a statistical test will correctly identify a significant increase in concentration above background. Note that power is typically defined with respect to a single test, not a network of tests. In this guidance, cumulative power is assessed for a single test over an *annual* period, depending on the frequency of the evaluation. Since some testing procedures may identify contamination more readily when several wells in the network are contaminated as opposed to just one or two, the Unified Guidance recommends that all testing strategies be compared on the following more stringent common basis.

The *effective power* of a testing protocol across a network of well-constituent pairs is defined as the probability of detecting contamination in the monitoring network when *one and only one* well-constituent pair is contaminated. Effective power is a conservative measure of how a testing regimen will perform across the network, because the set of statistical tests must uncover one contaminated well among many clean ones (*i.e.*, like 'finding a needle in a haystack'). As mentioned above, this initial judgment may need to be qualified with *effect size* and other site-specific considerations.

INTRODUCTION TO POWER CURVES

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 (**Figure 6-2**). 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 presented in units of standard deviations above the *baseline* (defined as the true background mean). This

allows different power curves to be compared across constituents, wells, or well-constituent pairs. These standardized units Δ in the case of normally-distributed data may be computed as:

$$\Delta = \frac{(\text{Mean Contamination Level}) - (\text{Mean Background Level})}{(\text{SD of Background Population})} \quad [6.1]$$

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 using Monte Carlo analysis on a computer. In these cases, power is typically estimated by generating normally-distributed random values at different mean contamination levels and repeatedly simulating the test procedure. With enough repetitions a reliable *power curve* can be plotted.

In the case of the normal power curve in **Figure 6-2**, the power values were computed analytically, using properties of the *non-central t-distribution.* In particular, the statistical power of a normal 99% prediction limit for the next single future value can be calculated as

$$1-\beta = \Pr\left\{T_{n-1}\left(\delta = \Delta\Big/\sqrt{1+\frac{1}{n}}\right) > t_{n-1,1-\alpha}\right\} \quad [6.2]$$

where Δ is the number of standardized (*i.e.,* standard deviation) units above the background population mean, (1–β) is the fractional power, δ is a non-centrality parameter, and:

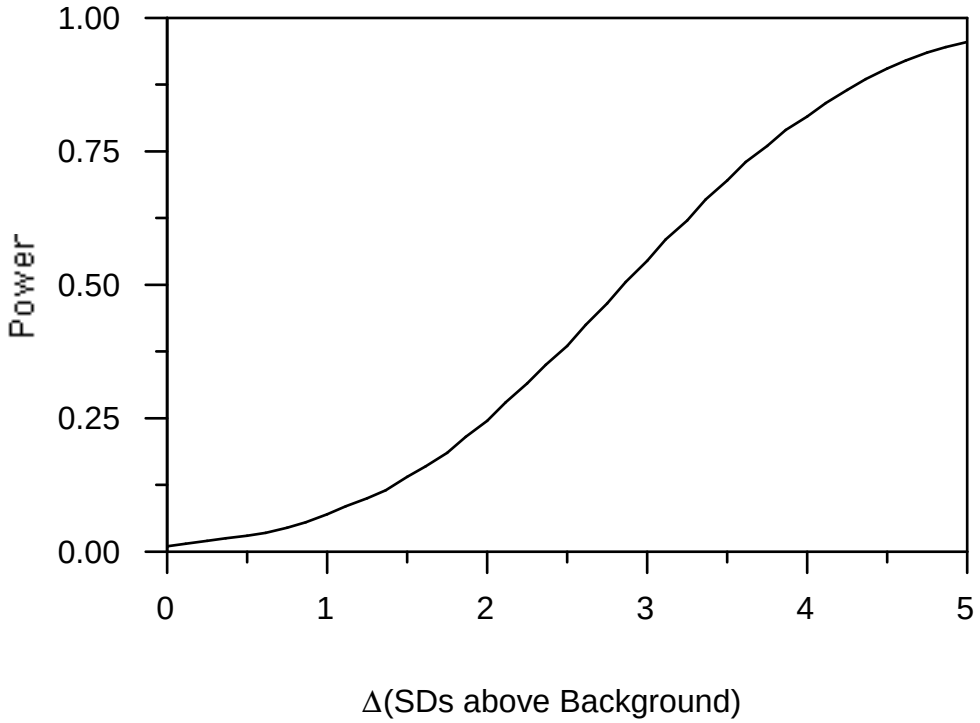
$$T_{n-1}\left(\delta = \Delta\Big/\sqrt{1+\frac{1}{n}}\right) \quad [6.3]$$

represents a non-central *t*-variate with (*n*–1) degrees of freedom and non-centrality parameter δ. Equation [6.2] was used with $n = 10$ to generate **Figure 6-2**.[7]

On a general power curve, the power at $\Delta = 0$ represents the false positive rate or *size* of the statistical test, because at that point no contamination is actually present (i.e., the background condition), even though the curve indicates how often a significant concentration increase will be detected. One should be careful to distinguish between the SWFPR across many statistical tests and the false positive rate represented on a curve measuring effective power. Since the effective power is defined as the testing procedure's ability to identify a *single* contaminated well-constituent pair, the effective power curve represents an *individual* test, *not* a network of tests. Therefore, the value of the curve at $\Delta = 0$ will only indicate the false positive rate associated with an individual test (α_{test}), not across the network as a whole. For many of the retesting strategies discussed in **Chapters 19** and **20**, the individual per-test false positive rate will be quite small and may appear to be nearly zero on the effective power curve.

[7] For users with access to statistical software containing the non-central T-distribution, this power curve can be duplicated. For example, the $\Delta = 3\sigma$ fractional power can be obtained using the following inputs: a central t-value of $t_{.99,\,9} = 2.821$, 9 df, and $\delta = 3\Big/\sqrt{1+(1/10)} = 2.8604$. The fractional power is .5414. It should be noted that the software may report the

Figure 6-2. Normal Power Curve (n = 10) for 99% Prediction Limit Test



To properly interpret a power curve, note that not only is the probability greater of identifying a concentration increase above background (shown as a decimal value between 0 and 1 along the vertical axis) as the magnitude of the increase gets bigger (as measured along the horizontal axis), but one can determine the probability of identifying certain kinds of increases. For instance, with effective power equivalent to that in **Figure 6-2**, any mean concentration increase of at least 2 background standard deviations will be detected about 25% percent of the time, while an increase of 3 standard deviations will be detected with approximately 55% probability or better than 50-50 odds. A mean increase of at least 4 standard deviations will be detected with about 80% probability.

An increase of 3 or 4 standard deviations above the baseline may or may not have practical implications for human health or the environment. That will ultimately depend on site-specific factors such as the constituents being monitored, the local hydrogeologic environment, proximity to environmentally sensitive populations, and the observed variability in background concentrations. In some circumstances, more sensitive testing procedures might be warranted. As a general guide especially in the absence of direct site-specific information, the Unified Guidance recommends that when background is approximately normal in distribution,[8] any statistical test should be able to detect a 3

probability as (β) rather than (1-β). For more complex power curves involving multiple repeat samples or multiple tests, integration is necessary to generate the power estimates.

[8] If a non-parametric test is performed, power (or more technically, efficiency) is often measured by Monte Carlo simulation using normally distributed data. So these recommendations also apply to that case.

standard deviation increase at least 55-60% of the time and a 4 standard deviation increase with at least 80-85% probability.

EPA REFERENCE POWER CURVES

Since effect sizes discussed in the next section often cannot or have not been quantified, the Unified Guidance recommends using the ERPC as a suitable basis of comparison for proposed testing procedures. Each reference power curve corresponds to one of three typical yearly statistical evaluation schedules — quarterly, semi-annual, or annual — and represents the cumulative power achievable during a single year at one well-constituent pair by a 99% upper (normal) prediction limit based on n = 10 background measurements and one new measurement from the compliance well (see **Chapter 18** for discussion of normal prediction limits). The ERPC are pictured in **Figure 6-3** below.

Any proposed statistical test procedure with effective power at least as high as the appropriate ERPC, especially in the range of three or more standard deviations above the background mean, should be considered to have reasonable power.[9] In particular, if the effective power first exceeds the ERPC at a mean concentration increase no greater than 3 background standard deviations (*i.e.*, $\Delta \leq 3$), the power is labeled 'good;' if the effective power first exceeds the ERPC at a mean increase between 3 and 4 standard deviations (*i.e.*, $3 < \Delta \leq 4$), the power is considered 'acceptable;' and if the first exceedance of the ERPC does not occur until an increase greater than 4 standard deviations (*i.e.*, $\Delta > 4$), the power is considered 'low.'

With respect to the ERPCs, one should keep the following considerations in mind:

1. The effective power of any testing method applied to a groundwater monitoring network can be increased merely by relaxing the SWFPR guideline, letting the SWFPR become larger than 10%. This is why a maximum annual SWFPR of 10% is suggested as standard guidance, to ensure fair power comparisons among competing tests and to limit the overall network-wide false positive rate.

2. The ERPCs are based on *cumulative* power over a one-year period. That is, if a single well-constituent pair is contaminated at standardized level Δ during each of the yearly evaluations, the ERPC indicates the probability that a 99% upper prediction limit test will identify the groundwater as impacted during at least one of those evaluations. Because the number of evaluations not only varies by facility, but also impacts the cumulative one-year power, different reference power curves should be employed depending on a facility's evaluation schedule. Quarterly evaluators should utilize the quarterly reference power curve (Q); semi-annual evaluators the semi-annual curve (S); and annual evaluators the annual curve (A).

3. If Monte Carlo simulations are used to evaluate the power of a proposed testing method, it should incorporate every aspect of the procedure, from initial screens of the data to final

[9] When using a retesting strategy in a larger network, the false positive rate associated with a single contaminated well (used to determine the effective power) will tend to be much smaller than the targeted SWFPR. Since the point at which the effective power curve intersects $\Delta = 0$ on the standardized horizontal axis represents the false positive rate for that individual test, the effective power curve by construction will almost always be *less* than the EPA reference power curve for small concentration increases above background. Of more concern is the relative behavior of the effective power curve at larger concentration increases, say two or more standard deviations above background.

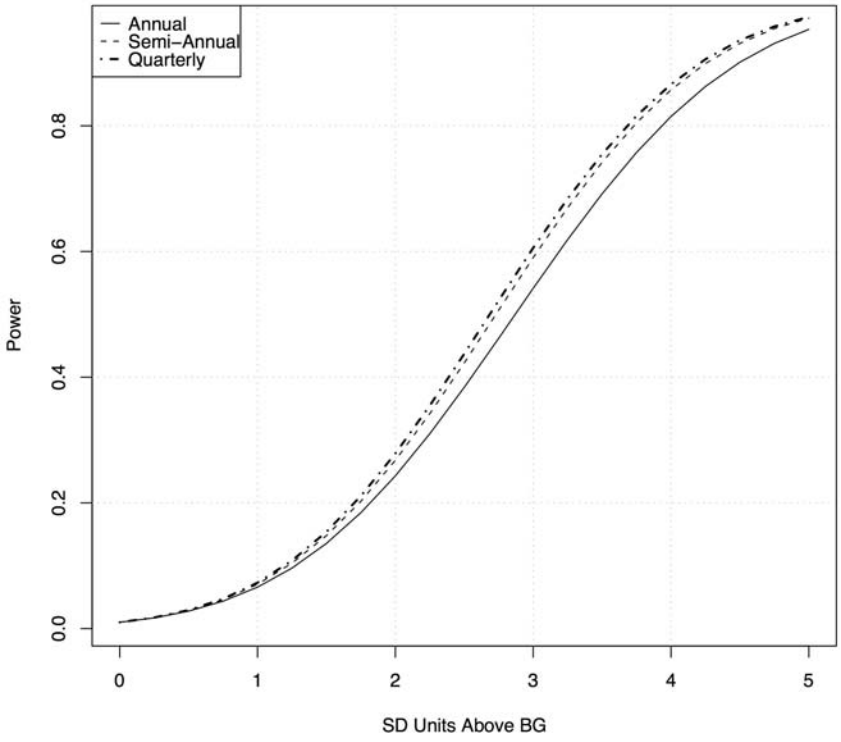
decisions concerning the presence of contamination. This is especially applicable to strategies that involve some form of retesting at potentially contaminated wells.

4. Although monitoring networks incorporate multiple well-constituent pairs, effective power can be gauged by simulating contamination in one and only one constituent at a single well.

5. The ERPCs should be considered a minimal power standard. The prediction limit test used to construct these reference curves does not incorporate retesting of any sort, and is based on evaluating a single new measurement from the contaminated well-constituent pair. In general, both retesting and/or the evaluation of multiple compliance point measurements tend to improve statistical power, so proposed tests that include such elements should be able to match the ERPC.

6. At sites employing multiple types of test procedures (***e.g.***, non-parametric prediction limits for some constituents, control charts for other constituents), effective power should be computed for each type of procedure to determine which type exhibits the least statistical power. Ensuring adequate power across the site implies that the ***least powerful*** procedure should match or exceed the appropriate ERPC, not just the most powerful procedure.

Figure 6-3. EPA Reference Power Curves



6.2.4 EFFECT SIZES AND DATA-BASED POWER CURVES

EFFECT SIZES

If site-specific or chemical-specific risk/health information is available particularly for naturally-occurring constituents, it can be used in some circumstances to develop an *effect size* of importance. An effect size (φ) is simply the smallest concentration increase above the mean background level that is presumed or known to have a measurable, deleterious impact on human health and/or the environment, or that would clearly signal the presence of contamination.

When an effect size can be quantified for a given constituent and is approved by the regulating authority, the acceptable power of the statistical test can be tailored to that amount. For instance, if an effect size for lead in groundwater at a particular site is φ = 10 ppb, one might require that the statistical procedure have an 80% or 95% chance of detecting such an increase. This would be true regardless of whether the power curve for lead at that site matches the ERPC. In some cases, an agreed-upon effect size will result in a more stringent power requirement compared to the ERPCs. In other cases, the power standard might be less stringent.

Effect sizes are not known or have not been determined for many groundwater constituents, including many inorganic parameters that have detection frequencies high enough to be amenable to effect size calculations. Because of this, many users will routinely utilize the relative power guidelines embodied in the ERPC. Even if a specific effect size cannot be determined, it is helpful to consider the site-specific and test-specific implications of a three or four standard deviation concentration increase above background. Taking the background sample mean ($\bar{x}$) as the estimated baseline, and estimating the underlying population variability by using the sample background standard deviation (s), one can compute the approximate actual concentrations associated with a three, four, five, *etc.* standard deviation increase above the baseline (as would be done in computing a *data-based power curve*; **Section 6.2.4**). These concentration values will only be approximate, since the true background mean (μ) and standard deviation (σ) are unknown. However, conducting this analysis can be useful in at least two ways. Each is illustrated by a simple example.

By associating the standardized units on a reference power curve with specific but approximate concentration levels, it is possible to evaluate whether the anticipated power characteristics of the chosen statistical method are adequate for the site in question. If not, another method with better power might be needed. ***Generally, it is useful to discuss and report statistical power in terms of concentration levels rather than theoretical units***.

►EXAMPLE 6-1

A potential permit GWPS for lead is 15 ppb, while natural background lead levels are normally distributed with an average of 6 ppb and a standard deviation of 2 ppb. The regulatory agency determines that a statistical test should be able to identify an exceedance of this GWPS with high power. Further assume that the power curve for a particular statistical test indicated 40% power at 3 standard deviations and 78% power at 4σ above background (a low power rating).

By comparing the actual standard deviation estimate to the required target increase φ = (15-6)/2 = 4.5 standard units, the power at the critical effect size would be 80% or higher using **Figure 6-2** as a

rough guide. This might be sufficient for monitoring needs even though the test did not meet the EPA reference criteria. Of course, the results apply only to this specific well-constituent test. ◄

For a given background sample, one can consider the regulatory and environmental impact of using *that particular background* as the basis of comparison in detection monitoring. Especially when deciding between *interwell* and *intrawell* tests at the same site, it is not unusual for the intrawell background from an individual well to exhibit much less variability than a larger set of observations pooled from multiple upgradient wells. This difference can be important since an intrawell test and an interwell test applied to the same site — *using identical relative power criteria* — might be associated with different risks to human health and the environment. A similar type of comparison might also aid in deciding whether the degrees of freedom of an intrawell test ought to be enlarged via a pooled estimate of the intrawell standard deviation (**Chapter 13**), whether a non-adjusted intrawell test is adequate, or whether more background sampling ought to be conducted prior to running intrawell tests.

►EXAMPLE 6-2

The standard deviation of an intrawell background population is σ_{intra} = 5 ppb, but that of upgradient, interwell background is σ_{inter} = 10 ppb. With the increased precision of an intrawell method, it may be possible to detect a 20 ppb increase with high probability (representing a $\Delta = 4\sigma_{intra}$ increase), while the corresponding probability for an interwell test is much lower (*i.e.*, 20 ppb = $2\sigma_{inter} = \Delta$). Of course, even if the intrawell test meets the ERPC target at four standardized units above background, consideration should be given as to whether or not 20 ppb is a meaningful increase.◄

One caveat is that calculation of either effect sizes or data-based power curves (see below) requires a reasonable estimate of the background standard deviation (σ). Such calculations may often be possible only for naturally-occurring inorganics or other constituents with fairly high detection frequencies in groundwater. Otherwise, power computations based on an effect size or the estimated standard deviation (s) are likely to be unreliable due to the presence of left-censored measurements (*i.e.*, non-detects).

A type of effect size calculation is presented in **Chapter 22** regarding methods for compliance/assessment and corrective action monitoring. A comparable effect size is computed by considering changes in mean concentration levels equal to a multiple of a fixed GWPS or clean-up/action level. While the mean level changes are multiples of the concentration limit and in that sense still relative, because they are tied to a fixed concentration standard, the power of the test can be linked to specific concentration levels.

DATA-BASED POWER CURVES

Even if basing power on a specific effect size is impractical for a given facility or constituent, it is still possible to relate power to absolute concentration levels rather than to the standardized units of the ERPC. While exact statistical power depends on the unknown population standard deviation (σ), an approximate power curve can be constructed based on the estimated background standard deviation (s). Instead of an estimate of power at a single effect size (depicted in **Example 6-1**), the actual power over a range of effect sizes can be evaluated. Such a graph is denoted in the Unified Guidance as a ***data-based power curve***, a term first coined by Davis (1998).

Since the sample standard deviation (*s*) is calculated from actual groundwater measurements, this in turn changes an abstract power curve based on relative concentrations (*i.e.*, $k\sigma$ units above the baseline mean) into one displaying approximate, but absolute, concentrations (*i.e.*, *ks* units above baseline). The advantages of this approach include the following:

- Approximate data-based power curves allow the user to determine statistical power at any desired effect size (ϕ).

- Even if the effect size (ϕ) is unspecified, data-based power curves tie the performance of the statistical test back to actual concentration levels of the population being tested.

- Once the *theoretical* power curve of a particular statistical test is known, a data-based power curve is extremely easy to construct. One merely substitutes the observed background standard deviation (*s*) for σ and multiply by *k* to determine concentration values along the horizontal axis of the power curve. Even if the theoretical power curve is unknown, the same calculations can be made on the reference curve to derive an approximate site-specific, data-based power curve for tests roughly matching the performance of the ERPCs.

- If the choice between an interwell test and an intrawell approach is a difficult one (**Section 6.3.2**), helpful power comparisons can be made between intrawell and interwell tests at the same site using data-based power curves. Even if both tests meet the ERPC criteria, they may be based on different sets of background measurements, implying that the *interwell* standard deviation (s_{inter}) might differ from the *intrawell* standard deviation (s_{intra}). By plotting both data-based power curves on the same set of axes, the comparative performance of the tests can be gauged.

►EXAMPLE 6-3

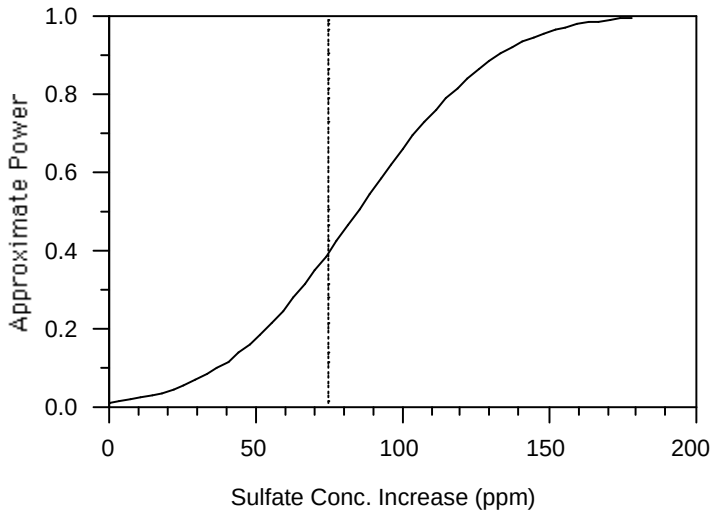
The following background sample is used to construct a test with theoretical statistical power similar to the ERPC for annual evaluations (see **Figure 6-2**). What will an approximate data-based power curve look like, and what is the approximate power for detecting a concentration increase of 75 ppm?

	Sulfate Concentrations (ppm)	
Quarter	BW-1	BW-2
1/95	560	550
4/95	530	570
7/95	568	540
10/95	490	542
1/96	510	590
Mean	545.0 ppm	
SD	29.7 ppm	

SOLUTION

The sample standard deviation of the pooled background sulfate concentrations is 29.7 ppm. Multiplying this amount by the number of standard deviations above background along the *x*-axis in **Figure 6-2** and re-plotting, the approximate data-based power curve of **Figure 6-3** can be generated. Then the statistical power for detecting an increase of 75 ppm is almost 40%.

Figure 6-3. Approximate *s*-Based Power Curve for Sulfate



Had the pooled sample size been n = 16 using the same test and sample statistics, a different and somewhat more powerful theoretical power curve would result. This theoretical curve can be generated (for a 1-of-1 prediction limit test) using the non-central T-distribution described earlier, if a user has the appropriate statistical software package. The power for a 75 ppm increase can be calculated using $\delta = 75/\sqrt{1+(1/16)} = 2.45$ and $t_{.99,\ 15}$ = 2.602, as closer to 46%. The larger background sample size makes for a more powerful test. ◄

6.2.5 SITES USING MORE THAN ONE STATISTICAL METHOD

There is no requirement that a facility apply one and only one statistical method to its groundwater monitoring program. The RCRA regulations explicitly allow for the use of multiple techniques, depending on the distributional properties of the constituents being monitored and the characteristics of the site. If some constituent data contain a high percentage of non-detect values, but others can be normalized, the statistical approach should vary by constituent.

With interwell testing, parametric prediction limits might be used with certain constituents and non-parametric prediction limits for other highly non-detect parameters. If intrawell testing is used, the most appropriate statistical technique for one constituent might differ at certain groups of wells than for others. Depending on the monitoring constituent, available individual well background, and other site-specific factors, some combination of intrawell prediction limits, control charts, and Wilcoxon rank-sum tests might come into play. At other sites, a mixture of intrawell and interwell tests might be conducted.

The Unified Guidance offers a range of possible methods which can be matched to the statistical characteristics of the observed data. The primary goal is that the statistical program should maximize the

odds of making correct judgments about groundwater quality. The guidance SWFPR and ERPC minimum power criteria serve as comprehensive guides for assessing any of the statistical methods.

One major concern is how statistical power should be compared when multiple methods are involved. Even if each method is so designed as not to exceed the recommended SWFPR, the effective power for identifying contaminated groundwater may vary considerably by technique and specific type of test. Depending on the well network and statistical characteristics of available data, a certain control chart test may or may not be as powerful as normal prediction limits. In turn, a specific non-parametric prediction limit test may be more powerful than some parametric versions. It is important that effective power be defined consistently, even at sites where more than one statistical method is employed.

The guidance encourages employing the *effective power* concept in assessing the ability of the statistical program to correctly identify and flag real concentration increases above background. As already defined, effective power is the probability that such an increase will be identified even if *only* one well-constituent pair is contaminated. Each well-constituent pair being tested should be considered equally at risk of containing a true increase above background. This also implies that the effective power of each statistical test in use should meet the criteria of the EPA reference curves. That is, the test with the *least* power should still have adequate power for identifying mean concentration increases.

The Unified Guidance does not recommend that *a single composite measure of effective power* be used to gauge a program's ability to identify potential contamination. To understand this last recommendation, consider the following hypothetical example. Two constituents exhibiting different subsurface travel times and diffusive potentials in the underlying aquifer are monitored with different statistical techniques. The constituent with the faster travel time might be measured using a test with very low effective power (compared to the ERPC), while the slower moving parameter is measured with a test having very high effective power. Averaging the separate power results into a single composite measure might result in an effective power roughly equivalent to the ERPC. Then the chances of identifying a release in a timely manner would be diminished unless rather large concentrations of the faster constituent began appearing in compliance wells. Smaller mean increases — even if 3 or 4 standard deviation units above background levels — would have little chance of being detected, while the time it took for more readily-identified levels of the slower constituent to arrive at compliance wells might be too long to be environmentally protective. Statistical power results should be reported separately, so that the effectiveness of each distinct test can be adequately judged. Further data-specific power evaluations could still be necessary to identify the appropriate test(s).

The following basic steps are recommended for assessing effective power at sites using multiple statistical methods:

1. Determine the number and assortment of distinct statistical tests. Different power characteristics may be exhibited by different statistical techniques. Specific control charts, *t*-tests, non-parametric prediction limits, *etc.* all tend to vary in their performance. The performance of a given technique is also strongly affected by the data characteristics. Background sample sizes, interwell versus intrawell choices, the number of retests and type of retesting plan, *etc.*, all affect statistical power. Each distinct data configuration and retesting plan will delineate a slightly different statistical test method.

2. Once the various methods have been identified, gauge the effective power of each.[10] Often the easiest way to measure power is via Monte Carlo simulation. Effective power involves a single well-constituent pair, so the simulation needs to incorporate only one population of background measurements representing the baseline condition and one population of compliance point measurements.

3. To run a Monte Carlo simulation, repeat the following algorithm a large number of times (*e.g.*, N = 10,000). Randomly generate a set of measurements from the background population in order to compute either a comparison limit for a control chart or some type of prediction limit test, or the background portion for a *t*-test or Wilcoxon rank-sum calculation, *etc*. Then generate compliance point samples at successively higher mean concentration levels, representing increases in standard deviation units above the baseline average. Perform each distinct test on the simulated data, recording the result of each iteration. By determining how frequently the concentration increase is identified at each successive mean level (including retests if necessary), the effective power for each distinct method can be estimated and compared.

►EXAMPLE 6-4

As a simple example of measuring effective power, consider a site using two different statistical methods. Assume that most of the constituents will be tested interwell with a 1-of-3 parametric normal prediction limit retesting plan for individual observations (**Chapter 19**). The remaining constituents having low detection rates and small well sample sizes will be tested intrawell with a Wilcoxon rank-sum test.

To measure the effective power of the normal prediction limits, note that the same number of background measurements (n = 30) is likely to be available for each of the relevant constituents. Since the per-constituent false positive rate (α_c) and the number of monitored wells (w) will also be identical for these chemicals, the same κ multiplier can be used for each prediction limit, despite the fact that the background mean and standard deviation will almost certainly vary by constituent.

Because of these identical data and well configurations, the effective power of each normal prediction limit will also be the same,[11] so that only one prediction limit test need be simulated. It is sufficient to assume the background population has a standard normal distribution. The compliance point population at the single contaminated well also has a normal distribution with the same standard deviation but a mean (μ) shifted upward to reflect successive relative concentration increases of 1 standard deviation, 2 standard deviations, 3 standard deviations, *etc*.

Simulate the power by conducting a large number of iterations (*e.g.*, N = 10,000-20,000) of the following algorithm: Generate 30 random observations from background and compute the sample mean

[10] Since power is a property of the statistical method and not linked to a specific data set, power curves are not needed for all well-constituent pairs, but only for each distinct statistical method. For instance, if intrawell prediction limits are employed to monitor barium at 10 compliance wells and the intrawell background sample size is the same for each well, only one power curve needs to be created for this group of tests.

[11] Statistical power measures the likely performance of the *technique* used to analyze the data, and is not a statement about the *data* themselves.

and standard deviation. Calculate the prediction limit by adding the background mean to κ times the background standard deviation. For a 1-of-3 retesting plan, generate 3 values from the compliance point distribution (*i.e.*, a normal distribution with unit standard deviation but mean equal to μ). If the first of these measurements does not exceed the prediction limit, record a score of zero and move on to the next iteration. If, however, the first value is an exceedance, test the second value and possibly the third. If either resample does not exceed the prediction limit, record a score of zero and move to the next iteration. But if both resamples are also exceedances, record a score of one. The fraction of iterations (*N*) with scores equal to one is an estimate of the effective power at a concentration level of μ standard deviations above the baseline.

In the case of the intrawell Wilcoxon rank-sum test, the power will depend on the number of intrawell background samples available at each well and for each constituent.[12] Assume for purposes of the example that all the intrawell background sizes are the same with $n = 6$ and that two new measurements will be collected at each well during the evaluation period. The power will also depend on the frequency of non-detects in the underlying groundwater population. To simulate this aspect of the distribution for each separate constituent, estimate the proportion (p) of observed non-detects across a series of wells. Then set a RL for purposes of the simulation equal to z_β, the pth quantile of the standard normal distribution.

Finally, simulate the effective power by repeating a large number of iterations of the following algorithm: Generate $n = 6$ samples from a standard normal distribution to represent intrawell background. Also generate two samples from a normal distribution with unit standard deviation and mean equal to μ to represent new compliance point measurements from a distribution with mean level equal to μ standard deviations above background. Classify any values as non-detects that fall below z_β. Then jointly rank the background and compliance values and compute the Wilcoxon rank-sum test statistic, making any necessary adjustments for ties (*e.g.*, the non-detects). If this test statistic exceeds its critical value, record a score of one for the iteration. If not, record a score of zero. Again estimate the effective power at mean concentration level μ as the proportion of iterations (*N*) with scores of one.

As a last step, examine the effective power for each of the two techniques. As long as the power curves of the normal prediction limit and the Wilcoxon rank-sum test ***both*** meet the criteria of the ERPCs, the statistical program taken as a whole should provide acceptable power. ◄

[12] Technically, since the Wilcoxon rank-sum test will often be applied to non-normal data, power will also depend fundamentally on the true underlying distribution at the compliance well. Since there may be no way to determine this distribution, approximate power is measured by assuming the underlying distribution is instead normal.

6.3 HOW KEY ASSUMPTIONS IMPACT STATISTICAL DESIGN

6.3.1 STATISTICAL INDEPENDENCE

IMPORTANCE OF INDEPENDENT, RANDOM MEASUREMENTS

Whether a facility is in detection monitoring, compliance/assessment, or corrective action, having an appropriate and valid sampling program is critical. All statistical procedures *infer* information about the underlying population from the observed sample measurements. Since these populations are only sampled a few times a year, observations should be carefully chosen to provide accurate information about the underlying population.

As discussed in **Chapter 3**, the mathematical theory behind standard statistical tests assumes that samples were *randomly* obtained from the underlying population. This is necessary to insure that the measurements are *independent* and *identically distributed* [i.i.d.]). Random sampling means that each possible concentration value in the population has an equal or known chance of being selected any time a measurement is taken. Only random sampling guarantees with sufficiently high probability that a set of measurements is adequately representative of the underlying population. It also ensures that human judgment will not bias the sample results, whether by intention or accident.

A number of factors make classical random sampling of groundwater virtually impossible. A typical small number of wells represent only a very small portion of an entire well-field. Wells are screened at specific depths and combine potentially different horizontal and vertical flow regimes. Only a minute portion of flow that passes a well is actually sampled. Sampling normally occurs at fixed schedules, not randomly.

Since a typical aquifer cannot be sampled at random, certain assumptions are made concerning the data from the available wells. It is first assumed that the selected well locations will generate concentration data similar to a randomly distributed set of wells. Secondly, it is assumed that groundwater flowing through the well screen(s) has a concentration distribution identical to the aquifer as a whole. This second assumption is unlikely to be valid unless groundwater is flowing through the aquifer at a pace fast enough and in such a way as to allow adequate mixing of the distinct water volumes over a relatively short (*e.g.*, every few months or so) period of time, so that groundwater concentrations seen at an existing well could also have been observed at other possible well locations.

Adequate sampling of aquifer concentration distributions cannot be accomplished unless enough time elapses between sampling events to allow different portions of the aquifer to pass through the well screen. Most closely-spaced sampling events will tend to exhibit a statistical dependence (*autocorrelation).* This means that pairs of consecutive measurements taken in a series will be positively correlated, exhibiting a stronger similarity in concentration levels than expected from pairs collected at random times. This would be particularly true for overall water quality indicators which are continuous throughout an aquifer and only vary slowly with time.

Another form of statistical dependence is *spatial correlation.* Groundwater concentrations of certain constituents exhibit natural spatial variability, *i.e.*, a distribution that varies depending on the location of the sampling coordinates. Spatially variable constituents exhibit mean and occasionally

variance differences from one well to another. Pairs of spatially variable measurements collected from the same or nearby locations exhibit greater similarity than those collected from distinct, widely-spaced, or distant wells.

Natural spatial variability can result from a number of geologic and hydrological processes, including varying soil composition across an aquifer. Various geochemical, diffusion, and adsorption processes may dominate depending on the specific locations being measured. Differential flow paths can also impact the spatial distribution of contaminants in groundwater, especially if there is limited mixing of distinct groundwater volumes over the period of sampling.

An adequate groundwater monitoring sampling program needs to account for not only site-specific factors such as hydrologic characteristics, projected flow rates, and directional patterns, but also meeting data assumptions such as independence. Statistical adjustments are necessary, such as selecting intrawell comparisons for spatially distinct wells or removing autocorrelation effects in the case of time dependence.

DARCY'S EQUATION AND AUTOCORRELATION

Past EPA guidance recommended the use of Darcy's equation as a means of establishing a minimum time interval between samples. When validly applied as a basic estimate of groundwater travel time in a given aquifer, the Darcy equation ensures that separate volumes of groundwater are being sampled (***i.e.***, physical independence). This increases the probability that the samples will also be statistically independent.

The Unified Guidance in **Chapter 14** also includes a discussion on applying Darcy's equation. Caution is advised in its use, however, since Darcy's equation *cannot guarantee* temporal independence. Groundwater travel time is only one factor that can influence the temporal pattern of aquifer constituents. The measurement process itself can affect time related dependency. An imprecise analytical method might impart enough additional variability to make the measurements essentially uncorrelated even in a short sampling interval. Changes in analytical methods or laboratories and even periodic re-calibration of analytical instrumentation can impart time-related dependencies in a data set regardless of the time intervals between samples.

The overriding interest is in the behavior of chemical contaminants in groundwater, not the groundwater itself. Many chemical compounds do not travel at the same velocity as groundwater. Chemical characteristics such as adsorptive potential, specific gravity, and molecular size can influence the way chemicals move in the subsurface. Large molecules, for example, will tend to travel slower than the average linear velocity of groundwater because of matrix interactions. Compounds that exhibit a strong adsorptive potential will undergo a similar fate, dramatically changing time of travel predictions using the Darcy equation. In some cases, chemical interaction with the matrix material will alter the matrix structure and its associated hydraulic conductivity and may result in an increase in contaminant mobility. This last effect has been observed, for instance, with certain organic solvents in clay units (see Brown and Andersen, 1981).

The Darcy equation is also not valid in turbulent and non-linear laminar flow regimes. Examples of these particular hydrological environments include karst and 'pseudo-karst' (***e.g.***, cavernous basalt and extensively fractured rock) formations. Specialized methods have been investigated by Quinlan (1989)

for developing alternative monitoring procedures. Dye tracing as described by Quinlan (1989) and Mull, *et al*. (1988) can be useful for identifying flow paths and travel times in these two particular environments; conventional groundwater monitoring wells are often of little value in designing an effective monitoring system in these type of environments.

Thus, we suggest that Darcy's equation not be exclusively relied upon to gauge statistical sampling frequency. At many sites, quarterly or semi-annual sampling often provides a reasonable balance between maintaining statistical independence among observations yet enabling early detection of groundwater problems. The Unified Guidance recommends three tools to explore or test for time-related dependence among groundwater measurements. Time series plots (**Chapter 9**) can be constructed on multiple wells to examine whether there is a time-related dependence in the pattern of concentrations. Parallel traces on such a plot may indicate correlation across wells as part of a natural temporal, seasonal or induced laboratory effect. For longer data series, direct estimates of the autocorrelation in a series of measurements from a single well can be made using either the *sample autocorrelation function* or the *rank von Neumann ratio* (**Section 14.2**).

DATA MIXTURES INCLUDING ALIQUOT REPLICATE SAMPLES

Some facility data sets may contain both single and aliquot replicate groundwater measurements such as duplicate splits. An entire data set may also consist of aliquot replicates from a number of independent water quality samples. The guidance recommends against using aliquot data directly in detection monitoring tests, since they are almost *never* statistically independent. Significant positive correlation almost always exists between such duplicate samples or among aliquot sets. However, it is still possible to utilize some of the aliquot information within a larger water quality data set.

Lab duplicates and field splits can provide valuable information about the level of measurement variability attributable to sampling and/or analytical techniques. However, to use them as separate observations in a prediction limit, control chart, analysis of variance [ANOVA] or other procedure, the test must be specially structured to account for multiple data values per sampling event.

Barring the use of these more complicated methods, one suggested strategy has been to simply average each set of field splits and lab duplicates and treat the resulting mean as a single observation in the overall data set. Despite eliminating the dependence between field splits and/or lab duplicates, *such averaging is not an ideal solution*. The variability in means of two correlated measurements is approximately 30% less than the variability associated with two single independent measurements. If a data set consists of a mixture of single measurements and lab duplicates and/or field splits, the variability of the averaged values will be less than the variability of the single measurements. This would imply that the final data set is not *identically* distributed.

When data are not identically distributed, the actual false positive and false negative rates of statistical tests may be higher or lower than expected. The effect of mixing single measurements and averaged aliquot replicates might be balanced out in a two-sample t-test if sample sizes are roughly equal. However, the impact of non-identically distributed data can be substantial for an upper prediction limit test of a future single sample where the background sample includes a mixture of aliquot replicates and single measurements. Background variability will be underestimated, resulting in a lowered prediction limit and a higher false positive rate.

One statistically defensible but expensive approach is to perform the same number of aliquot replicate measurements on all physical samples collected from background and compliance wells. Aliquot replicates can be averaged, and the same variance reduction will occur in all the final observations. The statistical test degrees of freedom, however, are based on the number of independent, averaged samples.

Mixing single and averaged aliquot data is a serious problem if the component of variability due to field sampling methods and laboratory measurement error is a substantial fraction of the overall sample variance. When natural variability in groundwater concentrations is the largest component, averaging aliquot replicate measurements will do little to weaken the assumption of identically-distributed data. Even when variability due to sampling and analytical methods is a large component of the total variance, if the percentage of samples with aliquot replicate measurements is fairly small (say, 10% or less), the impact of aliquot replicate averaging should usually be negligible. However, consultation with a professional statistician is recommended.

The simplest alternative is to randomly select one value from each aliquot replicate set along with all non-replicate individual measurements, for use in statistical testing. Either this approach or the averaged replicate method described above will result in smaller degrees of freedom than the strategy of using all the aliquots, and will more accurately reflect the statistical properties of the data.

CORRECTING FOR TEMPORAL CORRELATION

The Unified Guidance recommends two general methods to correct for observable temporal correlation. Darcy's equation is mentioned above as a rough guide to physical independence of consecutive groundwater observations. A more generally applicable strategy for yet-to-be-collected measurements involves adjusting the sampling frequency to avoid autocorrelation in consecutive sampling events. Where autocorrelation is a serious concern, the Unified Guidance recommends running a *pilot study* at two or three wells and analyzing the study data by using the sample autocorrelation function (**Section 14.3.1**). The autocorrelation function plots the strength of correlation between consecutive measurements against the time lag between sampling events. When the autocorrelation becomes insignificantly different from zero at a particular sampling interval, the corresponding sampling frequency is the maximum that will ensure uncorrelated sampling events.

Two other strategies are recommended for adjusting already collected data. First, a longer data series at a single well can be corrected for seasonality by estimating and removing the seasonal trend (**Section 14.3.3**). If both a linear trend *and* seasonal fluctuations are evident, the seasonal Mann-Kendall trend test can be run to identify the trend despite the seasonal effects (**Section 14.3.4**). A second strategy is for sites where a temporal effect (*e.g.*, temporal dependence, seasonality) is apparent across multiple wells. This involves estimating a temporal effect via a *one-way* ANOVA and then creating adjusted measurements using the ANOVA residuals (**Section 14.3.3**). The adjusted data can then be utilized in subsequent statistical procedures.

6.3.2 SPATIAL VARIATION: INTERWELL VS. INTRAWELL TESTING

ASSUMPTIONS IN BACKGROUND-TO-DOWNGRADIENT COMPARISONS

The RCRA groundwater monitoring regulations initially presume that detection monitoring background can be defined on the basis of a definable groundwater gradient. In a considerable number of situations, this approach is problematic. No groundwater gradient may be measurable for identifying upgradient and downgradient well locations around a regulated unit. The hydraulic gradient may change in direction, depth or magnitude due to seasonal fluctuations. Groundwater mounding or other flow anomalies can occur. At most locations, significant spatial variability among wells exists for certain constituents. Where spatial variation is a natural artifact of the site-specific geochemistry, differences between upgradient and downgradient wells are unrelated to on-site waste management practices.

Both the Subtitle C and Subtitle D RCRA regulations allow for a determination that background quality may include sampling of wells not hydraulically upgradient of the waste management area. The rules recognize that this can occur either when hydrological information is unable to indicate which wells are hydraulically upgradient or when sampling other wells will be "representative or more representative than that provided by the upgradient wells."

For upgradient-to-downgradient well comparisons, a crucial detection monitoring assumption is that downgradient well changes in groundwater quality are only caused by on-site waste management activity. Up- and down-gradient well measurements are also assumed to be comparable and equal on average unless some waste-related change occurs. If other factors trigger significant increases in downgradient well locations, it may be very difficult to pinpoint the monitored unit as the source or cause of the contaminated groundwater.

Several other critical assumptions apply to the interwell approach. It is assumed that the upgradient and downgradient well samples are drawn from the same aquifer and that wells are screened at essentially the same hydrostratigraphic position. At some sites, more than one aquifer underlies the waste site or landfill, separated by confining layers of clay or other less permeable material. The fate and transport characteristics of groundwater contaminants likely will differ in each aquifer, resulting in unique concentration patterns. Consequently, upgradient and downgradient observations may not be comparable (*i.e.*, drawn from the same statistical population).

Another assumption is that groundwater flows in a definable pathway from upgradient to downgradient wells beneath the regulated unit. If flow paths are incorrectly determined or this does not occur, statistical comparisons can be invalidated. For example, a real release may be occurring at a site known to have groundwater mounding beneath the monitored unit. Since the groundwater may move towards both the downgradient and upgradient wells, it may not be possible to detect the release if both sets of wells become equally or similarly contaminated. One exception to this might occur if certain analytes are shown to exhibit uniform behavior in both historical upgradient and downgradient wells (e.g., certain infrequently detected trace elements). As long as the flow pathway from the unit to the

downgradient wells is assured, then an interwell test based on this combined background could still reflect a real exceedance in the downgradient wells.[13]

Groundwater flow should also move at a sufficient velocity beneath the site, so that the same groundwater observed at upgradient well locations is subsequently monitored at downgradient wells in the course of an evaluation period (*e.g.*, six months or a year). If groundwater flow is much slower, measurements from upgradient and downgradient wells may be more akin to samples from two separate aquifers. Extraneous factors may separately influence the downgradient and background populations, confusing the determination of whether or not a release has occurred.

While statistical testing can determine whether there are significant differences between upgradient and downgradient well measurements, it cannot determine *why* such differences exist. That is primarily the concern of a hydrologist who has carefully reviewed site-specific factors. Downgradient concentrations may be greater than background because contamination of the underlying aquifer has occurred. The increase may be due to other factors, including spatially variable concentration levels attributable to changing soil composition and geochemistry from one well location to another. It could also be due to the migration of contaminants from off-site sources reaching downgradient wells. These and other factors (including those summarized in **Chapter 4** on SSI Increases) should be considered before deciding that statistically significant background-to-downgradient differences represent site-related contamination.

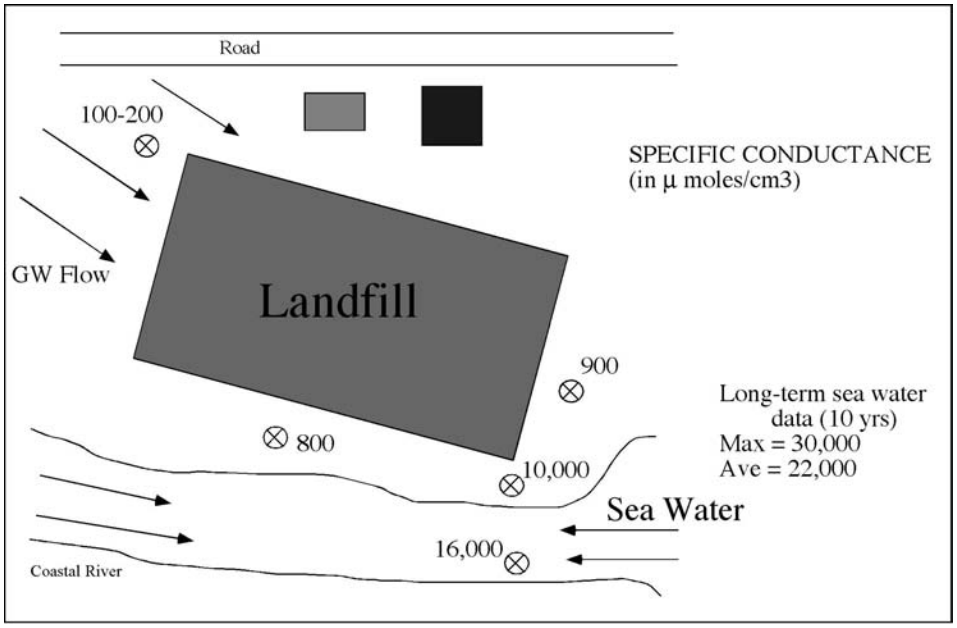
An example of how background-to-downgradient well differences can be misleading is illustrated in **Figure 6-4** below. At this Eastern coastal site, a Subtitle D landfill was located just off a coastal river emptying into the Atlantic Ocean a short distance downstream. Tests of specific conductance measurements comparing the single upgradient well to downgradient well data indicated significant increases at all downgradient wells, with one well indicating levels more than an order of magnitude higher than background concentrations.

Based on this analysis, it was initially concluded that waste management activities at the landfill had impacted groundwater. However, further hydrologic investigation showed that nearby river water also exhibited elevated levels of specific conductance, even higher than measurements at the downgradient wells. Tidal fluctuations and changes in river discharge caused sea water to periodically mix with the coastal river water at a location near the downgradient wells. Mixed river and sea water apparently seeped into the aquifer, impacting downgradient wells but not at the upgradient location. An off-site source as opposed to the landfill itself was likely responsible for the observed elevations in specific conductance. Without this additional hydrological information, the naive statistical comparison between upgradient and downgradient wells would have reached an incorrect conclusion.

[13] The same would be true of the "never-detected" constituent comparison, which does not depend on the overall flow pathway from upgradient to downgradient wells.

i

Figure 6-4. Landfill Site Configuration



TRADEOFFS IN INTERWELL AND INTRAWELL APPROACHES

The choice between interwell and intrawell testing primarily depends on the statistical characteristics of individual constituent data behavior in background wells. It is presumed that a thorough background study described in **Chapter 5** has been completed. This involves selecting the constituents deemed appropriate for detection monitoring, identifying distributional characteristics, and evaluating the constituent data for trends, stationarity, and mean spatial variability among wells. ANOVA tests can be used to assess both well mean spatial variability and the potential for pooled-variance estimates if an intrawell approach is needed.

As discussed in **Chapter 5**, certain classes of potential monitoring constituents are more likely to exhibit *spatial variation*. Water quality indicator parameters are quite frequently spatially variable. Some authors, notably Davis and McNichols (1994) and Gibbons (1994a), have suggested that significant spatial variation is a nearly ubiquitous feature at RCRA-regulated landfills and hazardous waste sites, thus invalidating the use of interwell test methods. The Unified Guidance accepts that interwell tests still have an important role in groundwater monitoring, particularly for certain classes of constituents like non-naturally occurring VOCs and some trace elements. Many sites may best be served by a statistical program which combines interwell and intrawell procedures.

Intrawell testing is an appropriate and recommended alternative strategy for many constituents. Well-specific backgrounds afford intrawell tests certain advantages over the interwell approach. One key advantage is *confounding results due to spatial variability* are eliminated, since all data used in an intrawell test are obtained from a single location. If natural background levels change substantially from

one well to the next, intrawell background provides the most accurate baseline for use in statistical comparisons.

At times, the ***variability*** in a set of upgradient background measurements pooled from multiple wells can be larger than the variation in individual intrawell background wells. Particularly if not checked with ANOVA well mean testing, interwell variability could substantially increase if changes in mean levels from one location to the next are also incorporated. While pooling should not occur among well means determined to be significantly different using ANOVA, a more likely situation is that pooled well true means and variance may be slightly different at each well. The ANOVA test might still conclude that the mean differences were insignificant and satisfy the equal variance assumption. The net result (as explained below) is that intrawell tests can be more statistically powerful than comparable interwell tests using upgradient background, despite employing a smaller background sample size.

Another advantage using intrawell background is that a ***reasonable baseline*** for tests of future observations can be established at historically contaminated wells. In this case, the intrawell background can be used to track the onset of even more extensive contamination in the future. Some compliance monitoring wells exhibit chronic elevated contaminant levels (e.g., arsenic) considerably above other site wells which may not be clearly attributed to a regulated unit release. The regulatory agency has the option of continuing detection monitoring or changing to compliance/corrective action monitoring. Unless the agency has already determined that the pre-existing contamination is subject to compliance monitoring or remedial action under RCRA, the detection monitoring option would be to test for recent or future concentration increases above the historical contamination levels by using intrawell background as a well-specific baseline.

Intrawell tests are not preferable for all groundwater monitoring scenarios. It may be unclear whether a given compliance well was historically contaminated prior to being regulated or more recently contaminated. Using intrawell background to set a baseline of comparison may ignore recent contamination subject to compliance testing and/or remedial action. Even more contamination in the future would then be required to trigger a statistically significant increase [SSI] using the intrawell test. The Unified Guidance recommends the use of intrawell testing only when it is clear that spatial variability is not the result of recent contamination attributable to the regulated unit.

A second concern is that intrawell tests typically utilize a smaller set of background data than interwell methods. Since statistical power depends significantly on background ***sample size***, it may be more difficult to achieve comparable statistical power with intrawell tests than with interwell methods. For the latter, background data can be collected from multiple wells when appropriate, forming a larger pool of measurements than would be available at a single well. However, it may also be possible to enhance intrawell sample sizes for parametric tests using the pooled- variance approach.

Traditional interwell tests can be appropriate for certain constituents if the hydraulic assumptions discussed earlier are verified and there is no evidence of significant spatial variability. Background data from other historical compliance wells not significantly different from upgradient wells using ANOVA may also be used in some cases. When these conditions are met, interwell tests can be preferable as generally more powerful tests. Upgradient groundwater quality can then be more easily monitored in

parallel to downgradient locations. Such upgradient monitoring can signal changes in natural in-situ concentrations or possible migration from off-site sources. [14]

For most situations, the background constituent data patterns will determine which option is most feasible. Clear indications of spatially distinct well means through ANOVA testing will necessitate some form of intrawell methods. Further choices are then which type of statistical testing will provide the best power.

It may be possible to increase the *effective sample size* associated with a series of intrawell tests. As explained in **Chapters 13 & 19**, the κ-multipliers for intrawell prediction limits primarily depend on the number of background measurements used to estimate the standard deviation. It is first necessary to determine that the intrawell background in a series of compliance wells is both uncontaminated and exhibits similar levels of variability from well to well. Background data from these wells can then be combined to form a *pooled* intrawell standard deviation estimate with larger degrees of freedom, even though individual well means vary. A transformation may be needed to stabilize the well-to-well variances. If one or more of the compliance wells is already contaminated, these should not be mixed with uncontaminated well data in obtaining the pooled standard deviation estimate.

A site-wide constituent pattern of no significant spatial variation will generally favor the interwell testing approach. But given the potential for hydrological and other issues discussed above, further evaluation of intrawell methods may be appropriate. **Example 6-2** provided an illustration of a specific intrawell constituent having a lower absolute standard deviation than an interwell pooled data set, and hence greater relative and absolute power. In making such an interwell-intrawell comparison, the specific test and all necessary design inputs must be considered. Even if a given intrawell data set has a low background standard deviation compared to an interwell counterpart, the advantage in absolute terms over the relative power approach will change with differing design inputs. The simplest way to determine if the intrawell approach might be advantageous is to calculate the actual background limits of a potential test using existing intra- and inter-well data sets. In a given prediction limit test, for example, the actual lower limit will determine the more powerful test.

If desired, approximate data-based power curves (**Section 6.2.4**) can be constructed to evaluate absolute power over a range of concentration level increases. In practice, the method for comparing interwell versus intrawell testing strategies with the same well-constituent pair involves the following basic steps:

1. Given the interwell background sample size (n_{inter}), the statistical test method (including any retesting), and the individual per-test α for that well-constituent pair, compute or simulate the relative power of the test at multiples of ks_{inter} above the baseline mean level. Let k range from 0 to 5 in increments of 0.5, where the interwell population standard deviation (σ_{inter}) has been replaced by the sample background standard deviation (s_{inter}).

[14] The same can be accomplished via intrawell methods if upgradient wells continue to be sampled along with required compliance well locations. Continued tracking of upgradient background groundwater quality is recommended regardless of the testing strategy.

2. Repeat Step 1 for the intrawell test. Use the intrawell background sample size (n_{intra}), statistical test method, background sample standard deviation (s_{intra}), and the same individual per-test α to generate a relative power curve.

3. On the same graph, plot overlays of the estimated data-based interwell and intrawell power curves (as discussed in **Section 6.2.4**). Use the same range of (absolute, not relative) concentration increases over baseline along the horizontal axis.

4. Visually inspect the data-based power curves to determine which method offers better power over a wider range of possible concentration increases.

The Unified Guidance recommends that users apply the most powerful statistical methods available in detecting and identifying contaminant releases for each well-constituent pair. The ERPC identifies a minimum acceptable standard for judging the relative power of particular tests. However, more powerful methods based on absolute power may be considered preferable in certain circumstances.

As a final concern, very small individual well samples in the early stages of a monitoring program may make it difficult to utilize an intrawell method having both sufficient statistical power and meeting false positive design criteria. One option would be to temporarily defer tests on those well-constituent pairs until additional background observations can be collected. A second option is to use the intrawell approach despite its inadequate power, until the intrawell background is sufficiently large via periodic updates (**Chapter 5**). A third option might be to use a more powerful intrawell test (e.g., a higher order 1-of-*m* parametric or non-parametric prediction limit test). Once background is increased, a lower order test might suffice. Depending on the type of tests, some control of power may be lost (parametric) or the false positive (non-parametric tests). These tradeoffs are considered more fully in **Chapter 19**. For the first two options and the parametric test under the third option, there is some added risk that a release occurring during the period of additional data collection might be missed. For the non-parametric test under the third option, there is an increased risk of a true false positive error. Any of these options might be included as special permit conditions.

6.3.3 OUTLIERS

Evaluation of outliers should begin with historical upgradient and possibly compliance well data considered for defining initial background, as described in **Chapter 5, Section 5.2.3**. The key goal is to select the data most representative of near-term and likely future background. Potentially discrepant or unusual values can occur for many reasons including 1) a contaminant release that significantly impacts measurements at compliance wells; 2) true but extreme background groundwater measurements, 3) inconsistent sampling or analytical chemistry methodology resulting in laboratory contamination or other anomalies; and 4) errors in the transcription of data values or decimal points. While the first two conditions may appear to be discrepant values, they would not be considered outliers.

When appraising extensive background data sets with long periods of acquisition and somewhat uncertain quality, it is recommended that a formal statistical evaluation of outliers not be conducted until a thorough review of data quality (errors, etc.) has been performed. Changes in analytical methodologies, the presence of sample interferences or dilutions can affect the historical data record. Past and current treatment of non-detects should also be investigated, including whether there are multiple reporting limits in the data base. Left-censored values can impact whether or not the sample

appears normal (**Chapter 15**), especially if the data need to be normalized via a transformation. Techniques for evaluating censored data should be considered, especially those which can properly account for multiple RLs. Censored probability plots (**Chapter 15**) or quasi-nonparametric box plots (**Chapter 12**) adapted by John Tukey (1977) can be used as methods to screen for outliers.

The guidance also recommends that statistical testing of potential outliers also be performed on initial background data, including historical compliance well data potentially considered as additional background data. Recognizing the potential risks as discussed in **Chapter 5**, removal of significant outliers may be appropriate even if no probable error or discrepancy can be firmly identified. The risk is that high values registering as statistical outliers may reflect an extreme, but real value from the background population rather than a true outlier, thereby increasing the likelihood of a false positive error. But the effect of removing outliers from the background data will usually be to improve the odds of detecting upward changes in concentration levels at compliance wells, and thus providing further protection of human health and the environment. *Automated screening and removal* of background data for statistical outliers is not recommended without some consideration of the likelihood of an outlier error.

A *statistical outlier* is defined as a value originating from a different statistical population than the rest of the sample. Outliers or observations not derived from the same population as the rest of the sample violate the basic statistical assumption of identically-distributed measurements. If an outlier is suspected, an initial helpful step is to construct a probability plot of the ordered sample data versus the standardized normal distribution (**Chapter 12**). A probability plot is designed to judge whether the sample data are consistent with a normal population model. If the data can be normalized, a probability plot of the transformed observations should also be constructed. Neither is a formal test, but can still provide important visual evidence as to whether the suspected outlier(s) should be further evaluated.

Formal testing for outliers should be done only if an observation seems particularly high compared to the rest of the sample. The data can be evaluated with either Dixon's or Rosner's tests (**Chapter 12**). These outlier tests assume that the rest of the data except for the suspect observation(s), are normally-distributed (Barnett and Lewis, 1994). It is recommended that tests also be conducted on transformed data, if the original data indicates one or more potential outliers. Lognormal and other skewed distributions can exhibit apparently elevated values in the original concentration domain, but still be statistically indistinguishable when normalized via a transformation. If the latter is the case, the outlier should be retained and the data set treated as fitting the transformed distribution.

Future background and compliance well data may also be periodically tested for outliers. However, removal of outliers should only take place under certain conditions, since a true elevated value may fit the pattern of a release or a change in historical background conditions. If either Dixon's or Rosner's test identifies an observation as a statistical outlier, the measurement should not be treated as such *until* a specific physical reason for the abnormal value can be determined. Valid reasons might include contaminated sampling equipment, laboratory contamination of the sample, errors in transcription of the data values, etc. Records documenting the sampling and analysis of the measurement (*i.e.*, the "chain of custody") should be thoroughly investigated. Based on this review, one of several actions might be taken as a general rule:

- If an error in transcription, dilution, analytical procedure, *etc.* can be identified and the correct value recovered, the observation should be replaced by its corrected value and further statistical analysis done with the corrected value.

- If it can shown that the observation is in error but the correct value cannot be determined, the observation should be removed from the data set and further statistical analysis performed on the reduced data set. The fact that the observation was removed and the reason for its removal should be documented when reporting results of the analysis.

- If no error in the value can be documented, it should be assumed that the observation is a true but extreme value. In this case, it should not be altered or removed. However, it may helpful to obtain another observation in order to verify or confirm the initial measurement.

6.3.4 NON-DETECTS

Statistically, non-detects are considered 'left-censored' measurements because the concentration of any non-detect is known or assumed only to fall within a certain range of concentration values (*e.g.*, between 0 and the RL). The direct estimate has been censored by limitations of the measurement process or analytical technique.

As noted, non-detect values can affect evaluations of potential outliers. Non-detects and detection frequency also impact what detection monitoring tests are appropriate for a given constituent. A low detection frequency makes it difficult to implement parametric statistical tests, since it may not be possible to determine if the underlying population is normal or can be normalized. Higher detection frequencies offer more options, including *simple substitution* or estimating the mean and standard deviation of samples containing non-detects by means of a *censored estimation technique* (**Chapter 15**).

Estimates of the background mean and standard deviation are needed to construct parametric prediction and control chart limits, as well as confidence intervals. If simple substitution is appropriate, imputed values for individual non-detects can be used as an alternate way to construct mean and standard deviation estimates. These estimates are also needed to update the *cumulative sum* [CUSUM] portion of control charts or to compute means of order *p* compared against prediction limits.

Simple substitution is not recommended in the Unified Guidance unless no more than 10-15% of the sample observations are non-detect. In those circumstances, substituting half the RL for each non-detect is not likely to substantially impact the results of statistical testing. Censored estimation techniques like *Kaplan-Meier* or *robust regression on order statistics* [ROS] are recommended any time the detection frequency is no less than 50% (see **Chapter 15)**.

For lower detection frequencies, non-parametric tests are recommended. Non-parametric prediction limits (**Chapter 18**) can be constructed as an alternative to parametric prediction limits or control charts. The Tarone-Ware two-sample test (**Chapter 16**) is specifically designed to accommodate non-detects and serves as an alternative to the *t*-test. By the same token, the Kruskal-Wallis test (**Chapter 17**) is a non-parametric, rank-based alternative to the parametric ANOVA. These latter tests can be used when the non-detects and detects can be jointly sorted and partially ordered (except for tied values).

When *all* data are non-detect, the Double Quantification rule (**Section 6.2.2**) can be used to define an approximate non-parametric prediction limit, with the RL as an upper bound. Before doing this, it should be determined whether chemicals never or not recently detected in groundwater should even be formally tested. This will depend on whether the monitored constituent from a large analytical suite is likely to originate in the waste or leachate.

Even if a data set contains only a small proportion of non-detects, care should be taken when choosing between the *method detection limit* [MDL], the quantification limit [QL], and the RL in characterizing 'non-detect' concentrations. Many non-detects are reported 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 estimate is reported, these samples represent 'detected, but not quantified' measurements. In this case, the actual concentration is assumed to be positive, falling somewhere between zero and the QL or possibly the RL.

Since the actual concentration is unknown, the suggested imputation when using simple substitution is to replace each non-detect having a qualifier of E or J by one-half the RL. Note, however, that E and J samples reported ***with*** estimated concentrations should be treated as valid measurements for statistical purposes. Substitution of one-half the RL is ***not recommended*** for these measurements, even though the degree of uncertainty associated with the estimated concentration is probably greater than that associated with measurements above the RL.

As a general rule, non-detect concentrations should ***not*** be assumed to be bounded above by the MDL. The MDL is usually estimated on the basis of ideal laboratory conditions with physical analyte samples that may or may not account for matrix or other interferences encountered when analyzing specific field samples. For certain trace element analytical methods, individual laboratories may report detectable limits closer to an MDL than a nominal QL. So long as the laboratory has confidence in the ability to quantify at its lab- or occasionally event-specific detection level, this RL may also be satisfactory. The RL should typically be taken as a more reasonable upper bound for non-detects when imputing estimated concentration values to these measurements.

RLs are sometimes but not always equivalent to a particular laboratory's QLs. While analytical techniques may change and improve over time leading to a lowering of the achievable QL, a contractually negotiated RL might be much higher. Often a multiplicative factor is built into the RL to protect a contract lab against particular liabilities. A good practice is to periodically review a given laboratory's capabilities and to encourage reporting non-detects with actual QLs whenever possible, and providing standard qualifiers with all data measurements as well as ***estimated*** concentrations for E- and J-flagged samples.

Even when no estimate of concentration can be made, a lab should regularly report the distinction between 'undetected' and 'detected, but not quantified' non-detect measurements. Data sets with such delineations can be used to advantage in rank-based non-parametric procedures. Rather than assigning the same tied rank to all non-detects (**Chapter 16**), 'detected but not quantified' measurements should be given larger ranks than those assigned to 'undetected' samples. These two types of non-detects should be treated as two *distinct* groups of tied observations for use in the non-parametric ***Wilcoxon rank-sum*** procedure.

6.4 DESIGNING DETECTION MONITORING TESTS

In the following sections, the main formal detection monitoring tests covered in this guidance are described in the context of site design choices. Advantages as well as limitations are presented, including the use of certain methods as diagnostic tools in determining the appropriate formal test(s).

6.4.1 T-TESTS

A statistical comparison between two sets of data is known as a two-sample test. When normality of the sample data can be presumed, the parametric Student *t*-test is commonly used (**Section 16.1**). This test compares two distinct populations, represented by two *samples*. These samples can either be individual well data sets, or a common pooled background versus individual compliance well data. The basic goal of the *t*-test is to determine whether there is any statistically significant difference between the two population means. Regulatory requirements for formal use of two-sample t-tests are limited to the Part 265 indicator parameters, and have generally been superseded in the Parts 264 and 258 rules by tests which can account for multiple comparisons.

When the sample data are non-normal and may contain non-detects, the Unified Guidance provides alternative two-sample tests to the parametric t-test. The Wilcoxon rank-sum test (**Section 16.2**) requires that the combined samples be sorted and ranked. This test evaluates potential differences in population *medians* rather than the *means*. The Tarone-Ware test (**Section 16.3**) is specially adapted to handle left-censored measurements, and also tests for differences in population medians.

The *t*-test or a non-parametric variant is recommended as a validation tool when updating intrawell or other background data sets (**Chapter 5**). More recently collected data considered for background addition are compared to the historical data set. A non-significant test result implies no mean differences, and the newer data may be added to the original set. These tests are generally useful for any two-sample diagnostic comparisons.

6.4.2 ANALYSIS OF VARIANCE [ANOVA]

The parametric one-way ANOVA is an extension of the t-test to multiple sample groups. Like its two-sample counterpart, ANOVA tests for significant differences in one or more group (e.g., well) means. If an overall significant difference is found as measured by the F-statistic, *post-hoc* statistical contrasts may be used to determine where the differences lie among individual group means. In the groundwater detection monitoring context, only differences of mean well increases relative to background are considered of importance. The ANOVA test also has wide applicability as a diagnostic tool.

USE OF ANOVA IN FORMAL DETECTION MONITORING TESTS

RCRA regulations under Parts 264 and 258 identify parametric and non-parametric ANOVA as potential detection monitoring tests. Because of its flexibility and power, ANOVA can sometimes be an appropriate method of statistical analysis when groundwater monitoring is based on an *interwell* comparison of background and compliance well data. Two types of ANOVA are presented in the Unified Guidance: parametric and non-parametric one-way ANOVA (**Section 17.1**). Both methods

attempt to assess whether distinct monitoring wells differ in average concentration during a given evaluation period.[15]

Despite the potential attractiveness of ANOVA tests, use in formal detection monitoring is limited by these important factors:

- ❖ Many monitoring constituents exhibit significant spatial variability and cannot make use of interwell comparisons;
- ❖ The test can be confounded by a large number of well network comparisons;
- ❖ A minimum well sample size must be available for testing; and
- ❖ Regulatory false positive error rate restrictions limit the ability to effectively control the overall false positive rate.

As discussed in **Section 6.2.3**, many if not most inorganic monitoring constituents exhibit spatial variability, precluding an interwell form of testing. Since ANOVA is inherently an interwell procedure, the guidance recommends against its use for these constituents and conditions. Spatial variability implies that the average groundwater concentration levels vary from well to well because of existing on-site conditions. Mean differences of this sort can be identified by ANOVA, but the cause of the differences cannot. Therefore, results of a statistically significant ANOVA might be falsely attributed as a regulated unit release to groundwater.

ANOVA testing might be applied to synthetic organic and trace element constituent data. However, spatial variation across a site is also likely to occur from offsite or prior site-related organic releases. An existing contamination plume generally exhibits varying average concentrations longitudinally, as well as in cross-section and depth. For other organic constituents never detected at a site, ANOVA testing would be unnecessary. Certain trace elements like barium, arsenic and selenium do often exhibit some spatial variability. Other trace element data generally have low overall detection rates, which may also preclude ANOVA applications. Overall, very few routine monitoring constituents are measurable (i.e., mostly detectable) yet not spatially distinct to warrant using ANOVA as a formal detection monitoring test. Other guidance tests better serve this purpose.

ANOVA has good power for detecting real contamination provided the network is small to moderate in size. But for large monitoring networks, it may be difficult to identify single well contamination. One explanation is that the ANOVA F-statistic simultaneously combines all compliance well effects into a single number, so that many other uncontaminated wells with their own variability can mask the test effectiveness to detect the contaminated well. This might occur at larger sites with multiple waste units, or if only the edge of a plume happens to intersect one or two boundary wells.

The statistical power of ANOVA depends significantly on having at least 4 observations per well available for testing. Since the measurements must be statistically independent, collection of four well observations may necessitate a wait of several months to a few years if the natural groundwater velocity

[15] Parametric ANOVA assesses differences in means; the non-parametric ANOVA compares ***median*** concentration levels. Both statistical measures are a kind of average.

is low. In this case, other strategies (e.g., prediction limits) might be considered that allow each new groundwater measurement to be tested as it is collected and analyzed.

The one-way ANOVA test in the RCRA regulations is not designed to control the false positive error rate for multiple constituents. The rules mandate a minimum false positive error rate (α) of 5% per test application. With an overall false positive rate of approximately 5% per constituent, a potentially very high SWFPR can result as the number of constituents tested by ANOVA increases and if tests are conducted more than once per year.

For these reasons, the Unified Guidance does not generally recommend ANOVA for formal detection monitoring. ANOVA might be applicable to a small number of constituents, depending on the site. Prediction limit and control chart strategies using retesting are usually more flexible and offer the ability to accommodate even very large monitoring networks, while meeting the false positive and statistical power targets recommended by the guidance.

USE OF ANOVA IN DIAGNOSTIC TESTING

In contrast, ANOVA is a versatile tool for diagnostic testing, and is frequently used in the guidance for that purpose. Parametric or non-parametric one-way versions are the principal means of identifying prior spatial variability among background monitoring wells (**Chapter 13**). Improving sample sizes using intrawell pooled variances also makes use of ANOVA (**Chapter 13**). Equality of variances among wells is evaluated with ANOVA (**Chapter 11**). ANOVA is also applied when determining certain temporal trends in parallel well sample constituent data (**Chapter 14**).

Tests of natural spatial variability can be made by running ANOVA prior to any waste disposal at a new facility located above an undisturbed aquifer (Gibbons, 1994a). If ANOVA identifies significant upgradient and downgradient well differences when wastes have not yet been managed on-site, natural spatial variability is the likely cause. Prior on-site contamination might also be revealed in the form of significant ANOVA differences.

Sites with multiple upgradient background wells can initially conduct an ANOVA on historical data from just these locations. Where upgradient wells are not significantly different for a given constituent, ANOVA testing can be extended to existing historical compliance well data for evaluating potential additions to the upgradient background data base.

If intrawell tests are chosen because of natural spatial variation, the results of a one-way ANOVA on background data from multiple wells can sometimes be used to improve intrawell background limits (**Section 13.3**). Though the amount of intrawell background at any given well may be small, the ANOVA provides an estimate of the *root mean squared error* [RMSE], which is very close to an estimate of the *average per-well standard deviation*. By substituting the RMSE for the usual well-specific standard deviation (s), a more powerful and accurate intrawell limit can be constructed, at least at those sites where intrawell background across the group of wells can be normalized and the variances approximately equalized using a common transformation.

Although the Unified Guidance primarily makes use of one-way ANOVA, many kinds of ANOVA exist. The one-way ANOVA applications so far discussed— in formal detection monitoring or to assess well mean differences— utilize data from spatial locations as the factor of interest. In some situations,

correlated behavior may exist for a constituent among well samples evaluated in different temporal events. A constituent measured in a group of wells may simultaneously rise or fall in different time periods. Under these conditions, the data are no longer random and independent. ANOVA can be used to assess the significance of such systematic changes, making *time* the factor of interest. Time can also play a role if the sample data exhibit cyclical seasonal patterns or if parallel upward or downward trends are observed both in background and compliance point wells.

If time is an important second factor, a *two-way* ANOVA is probably appropriate. This procedure is discussed in Davis (1994). Such a method can be used to test for and adjust data either for seasonality, parallel trends, or changes in lab performance that cause temporal (*i.e.*, time-related) effects. It is somewhat more complicated to apply than a one-way test. The main advantage of a two-way ANOVA is to separate components of overall data variation into three sources: well-to-well mean-level differences, temporal effects, and random variation or statistical error. Distinguishing the sources of variation provides a more powerful test of whether significant well-to-well differences actually exist compared to using only a one-way procedure.

A significant temporal factor does not necessarily mean that the one-way ANOVA will *not* identify actual well-to-well spatial differences. It merely does not have as strong a chance of doing so. Rarely will the one-way ANOVA identify non-existent well-to-well differences. One situation where this can occur is when there is a strong *statistical interaction* between the well-to-well factor and the time factor in the two-way ANOVA. This would imply that changes in lab performance or seasonal cycles affect certain wells (*e.g.*, compliance point) to a different degree or in a different manner than other wells (*e.g.*, background). If this is the case, professional consultation is recommended before conducting more definitive statistical analyses.

6.4.3 TREND TESTS

Most formal detection monitoring tests in the guidance compare background and compliance point populations under the key assumption that the populations *are stationary over time*. The distributions in each group or well are assumed to be stable during the period of monitoring, with only random fluctuations around a constant mean level. If a significant trend occurs in the background data, these tests cannot be directly used. Trends can occur for several reasons including natural cycles, gradual changes in aquifer parameters or the effects of contaminant migration from off-site sources.

Although not specifically provided for in the RCRA regulations, the guidance necessarily includes a number of tests for evaluating potential trends. **Chapter 17, Section 17.3** covers three basic trend tests. (1) *Linear regression* is a parametric method requiring normal and independent trend residuals, and can be used both to identify a linear trend and estimate its magnitude; (2) For non-normal data (including sample data with left-censored measurements), the *Mann-Kendall* test offers a non-parametric method for identifying trends; and (3) To gauge trend magnitude with non-normal data, the *Theil-Sen* trend line can be used.

Trend analyses are primarily diagnostic tests, which should be applied to background data prior to implementing formal detection monitoring tests. If a significant trend is uncovered, two options may apply. The particular monitoring constituent may be dropped in favor of alternate constituents not exhibiting non-stationary behavior. Alternatively, prediction limit or control chart testing can make use of stationary *trend residuals* for testing purposes. One limitation of the latter approach requires making

an assumption that the historical trend will continue into future monitoring periods. In addition, future data needs to be de-trended prior to testing. If a trend happened to be of limited duration, this assumption may not be reasonable and could result in identifying a background exceedance when it does not exist. If a trend occurs in future data at a compliance well and prior background data was stationary, other detection monitoring tests are likely to eventually identify it. Trend testing may also be applied to once-future data considered for a periodic background update, although the guidance primarily relies on t-testing of historical and future groups to assess data suitability.

At historically contaminated compliance wells, establishing a proper baseline for a prediction limit or control chart is problematic, since uncontaminated concentration data cannot be collected. Depending on the pattern of contamination, an intrawell background may either have a stable mean concentration level or exhibit an increasing or decreasing trend. Particularly when intrawell background concentrations are rising, the assumption of a static baseline population required by prediction limits and control charts will be violated.

As an alternative, the Unified Guidance recommends a test for trend to measure the extent and nature of the apparent increase. Trend testing can determine if there is a statistically significant positive trend over the period of monitoring and can also determine the magnitude (*i.e.*, slope) of the trend. In identifying a positive trend, it might be possible to demonstrate that the level of contamination has increased relative to historical behavior and indicate how rapidly levels are increasing.

Trend analyses can be used directly as an alternative test against a GWPS in compliance and corrective action monitoring. For typical compliance monitoring, data collected at each compliance well are used to generate a lower confidence limit compared to the fixed standard (**Chapters 7, 21 and 22**). A similar situation occurs when corrective action is triggered, but making use of an upper confidence interval for comparison. For compliance well data containing a trend, the appropriate confidence interval is constructed around a linear regression trend line (or its non-parametric alternative) in order to better estimate the most current concentration levels. Instead of a single confidence limit for stationary tests, the confidence limit (or band) estimate changes with time.

6.4.4 STATISTICAL INTERVALS

Prediction limits, tolerance limits, control chart limits and confidence limits belong to the class of methods known as statistical intervals. The first three are used to define their respective detection monitoring test limits, while the last is used in fixed standard compliance and corrective action tests. When using a background GWPS, either approach is possible (see **Section 7.5**). Intervals are generated as a statistic from reference sample data, and represent a probable range of occurrence either for a future sample statistic or some parameter of the population (in the case of confidence intervals) from which the sample was drawn. A future sample statistic might be one or more single values, as well as a future mean or median of specific size, drawn from one or more sample sets to be compared with the interval (generally an upper limit). Both the reference and comparison sample populations are themselves unknown, with the latter initially presumed to be identical to the reference set population. In the groundwater monitoring context, the initial reference sample is the background data set.

The key difference in confidence limits[16] is that a statistical interval based on a single sample is used to estimate the probable range of a population parameter like the true mean, median or variance. The three detection monitoring tests use intervals to identify ranges of future sample statistics likely to arise from the background population based on the initial sample, and are hence two- or multiple-sample tests.

Statistical intervals are inherently two-sided, since they represent a finite range in which the desired statistic or population parameter is expected to occur. Formally, an interval is associated with a level of confidence (1-α); by construction, the error rate α represents the remaining likelihood that the interval *does not contain* the appropriate statistic or parameter. In a two-sided interval, the α-probability is associated with ranges both above and below the statistical interval. A one-sided upper interval is designed to contain the desired statistic or parameter at the same (1-α) level of confidence, but the remaining error represents only the range above the limit. As a general rule, detection monitoring options discussed below use one-sided upper limits because of the nature of the test hypotheses.

PREDICTION LIMITS

Upper prediction limits (or intervals) are constructed to contain with (1-α) probability, the next few sample value(s) or sample statistic(s) such as a mean from a background population. Prediction limits are exceptionally versatile, since they can be designed to accommodate a wide variety of potential site monitoring conditions. They have been extensively researched, and provide a straightforward interpretation of the test results. Since this guidance strongly encourages use of a comprehensive design strategy to account for both the cumulative SWFPR and effective power to identify real exceedances, prediction limit options offer a most effective means of accounting for both criteria. The guidance provides test options in the form of parametric normal and non-parametric prediction limit methods. Since a retesting strategy of some form is usually necessary to meet both criteria, prediction limit options are constructed to formally include resampling as part of the overall tests.

Chapters 18 and 19 provide nine parametric normal prediction limit test options: four tests of future values (1-of-2, 1-of-3, 1-of-4 or a modified California plan) and five future mean options (1-of-1, 1-of-2, or 1-of-3 tests of mean size 2, and 1-of-1 or 1-of-2 tests of mean size 3). Non-parametric prediction limit options cover the same future value test options as the parametric versions, as well as two median tests of size 3 (1-of-1 or 1-of-2 tests). **Appendix D** tables provide the relevant κ-factors for each parametric normal test option, the achievable false positive rates for non-parametric tests, and a categorical rating of relative test power for each set of input conditions. Prediction limits can be used both for interwell and intrawell testing. Selecting from among these options should allow the two site design criteria to be addressed for most groundwater site conditions.

The options provided in the guidance are based on a wider class known in the statistical literature as *p-of-m* prediction limit tests. Except for the two modified California plan options, those selected are 1-of-*m* test varieties. The number of future measurements to be predicted (i.e., contained) by the interval is also denoted in the Unified Guidance by *m* and can be as small as $m = 1$. To test for a release to groundwater, compliance well measurements are designated as future observations. Then a limit is constructed on the background sample, with the prediction limit formula based on the number of *m*

[16] Confidence limits are further discussed **in Chapters 7, 21 and 22** for use in compliance and corrective action testing.

future values or statistics to be tested against the limit. As long as the compliance point measurements are similar to background, the prediction limit should contain all *m* of the future values or statistics with high probability (the level of confidence). For a 1-of-*m* test, all *m* values must be larger than the prediction limit to be declared an exceedance, as initial evidence that compliance point concentrations are higher than background.

Prediction limits with retesting are presented in **Chapter 19**. When retesting is part of the procedure, there are significant and instructive differences in statistical performance between parametric and non-parametric prediction limits.

Parametric prediction limits are constructed using the general formula: $PL = \bar{x} + \kappa \cdot s$, where $\bar{x}$ and *s* are the background sample mean and standard deviation, and κ is the specific multiplicative factor for the type of test, background sample size, and the number of annual tests. The number of tests made against a common background is also an input factor for interwell comparison. The **Appendix D** κ-factors are specifically designed to meet the SWFPR objective, but power will vary. Larger background sample sizes and higher order (*m*) tests afford greater power.

When background data cannot be normalized, a non-parametric prediction limit can be used instead. A non-parametric prediction limit test makes use of one or another of the largest sample values from the background data set as the limit. For a given background sample size and test type, the level of confidence of that maximal value is fixed.

Using the absolute maximum of a background data set affords the highest confidence and lowest single-test false positive error. However, even this confidence level may not be adequate to meet the SWFPR objective, especially for lower order 1-of-*m* tests. A higher order future values test using the same maximum and background sample size will provide greater false positive confidence and hence a lower false positive error rate. For a fixed background sample size, a 1-of-4 retesting scheme will have a lower achievable significance level (α) than a 1-of-3 or 1-of-2 plan for any specific maximal value. A larger background sample size using a fixed maximal value for any test also has a higher confidence level (lower α) than a smaller sample.

But for a fixed non-parametric limit of a given background sample size, the power decreases as the test order increases. If the non-parametric prediction limit is set at the maximum, a 1-of-2 plan will be more powerful than a 1-of-4 plan. It is relatively easy to understand why this is the case. A verified exceedance in a 1-of-2 test occurs only if two values exceed the limit, but would require four to exceed for the 1-of-4 plan. As a rule, even the highest order non-parametric test using some maximal background value will be powerful enough to meet the ERPC power criteria, but achieving a sufficiently low single-test error rate to meet the SWFPR is more problematic.

If the SWFPR objective can be attained at a maximum value for higher order 1-of-*m* tests, it may be possible to utilize lower maxima from a large background data base. Lower maxima will have greater power and a somewhat higher false positive rate. Limited comparisons of this type can be made when choosing between the largest or second-largest order statistics in the Unified Guidance **Appendix D Tables 19-19** to **19-24**. A more useful and flexible comparison for 1-of-*m* future value plans can be obtained using the EPA Region 8 *Optimal Rank Values Calculator* discussed in **Chapter 19**. The calculator identifies the lowest ranked maximal value of a background data set for 1-of-1 to 1-of-4 future

value non-parametric tests which can meet the SWFPR objective, while providing ERPC ratings and fractional power estimates at 2, 3, and 4 standard deviations above background.

TOLERANCE INTERVALS

Tolerance intervals are presented in **Section 17.2**. A tolerance interval is generated from background sample data to contain a pre-specified *proportion* of the underlying population (*e.g.*, 99% of all possible population measurements) at a certain level of confidence. Measurements falling outside the tolerance interval can be judged to be statistically different from background.

While tolerance intervals are an acceptable statistical technique under RCRA as discussed in **Section 2.3**, the Unified Guidance generally recommends prediction limits instead. Both methods can be used to compare compliance point measurements to background in detection monitoring. The same general formula is used in both tests for constructing a parametric upper limit of comparison: $\bar{x} + \kappa s$. For non-parametric upper limit tests, both prediction limits and tolerance intervals use an observed order statistic in background (often the background maximum). But prediction limits are ultimately more flexible and easier to interpret than tolerance intervals.

Consider a parametric upper prediction limit test for the next two compliance point measurements with 95% confidence. If either measurement exceeds the limit, one of two conditions is true: either the compliance point distribution is significantly different and higher than background, or a false positive has been observed and the two distributions are similar. False positives in this case are expected to occur 5% of the time. Using an upper tolerance interval is not so straightforward. The tolerance interval has an extra statistical parameter that must be specified — the coverage (γ) — representing the fraction of background to be contained beneath the upper limit. Since the confidence level (1–α) governs how often a statistical interval contains its target population parameter (**Section 7.4**), the complement α does not necessarily represent the false positive rate in this case.

In fact, a tolerance interval constructed with 95% confidence to cover 80% of background is designed so that as many as 20% of all background measurements will exceed the limit with 95% probability. Here, α = 5% represents the probability that the true coverage will be less than 80%. But less clear is the false positive rate of a tolerance interval test in which as many as 1 in 5 background measurements are expected to exceed the upper background limit. Are compliance point values above the tolerance interval indicative of contaminated groundwater or merely representative of the upper ranges of background?

Besides a more confusing interpretation, there is an added concern. Mathematically valid retesting strategies can be computed for prediction limits, but not yet for tolerance intervals, further limiting their usefulness in groundwater testing. It is also difficult to construct powerful *intrawell* tolerance intervals, especially when the intrawell background sample size is small. Overall, there is little practical need for two similar (but not identical) methods in the Unified Guidance, at least in detection monitoring.

If tolerance intervals ***are*** employed as an alternative to *t*-tests or ANOVA when performing interwell tests, the RCRA regulations allow substantial flexibility in the choice of α. This means that a somewhat arbitrarily high confidence level (1–α) can be specified when constructing a tolerance interval. However, unless the coverage coefficient (γ) is also set to a high value (*e.g.*, ≥ 95%), the test is likely to incur a large risk of false positives despite a small α.

One setting in which an upper tolerance interval is very appropriate is discussed in **Section 7.5**. Some constituents that must be evaluated under compliance/assessment or corrective action may not have a fixed GWPS. Existing background levels may also exceed a fixed GWPS. In these cases, a background standard can be constructed using an upper tolerance interval on background with 95% confidence and 95% coverage. The standard will then represent a reasonable upper bound on background and an achievable target for compliance and remediation testing.

6.4.5 CONTROL CHARTS

Control charts (**Chapter 20**) are a viable alternative to prediction limits in detection monitoring. One advantage of a control chart over a prediction limit is that control charts allow compliance point data to be viewed and assessed graphically over time. Trends and changes in concentration levels can be easily seen, because the compliance measurements are consecutively plotted on the chart as they are collected, giving the data analyst an historical overview of the concentration pattern. Standard prediction limits allow only *point-in-time comparisons* between the most recent data and background, making long-term trends more difficult to identify.

The guidance recommends use of the combined *Shewhart-CUSUM control chart.* The advantage is that *two* statistical quantities are assessed at every sampling event, both the new individual measurement and the cumulative sum [CUSUM] of past and current measurements. Prediction limits do not incorporate a CUSUM, and this can give control charts comparatively greater sensitivity to gradual (upward) trends and shifts in concentration levels. To enhance false positive error rate control and power, retesting can also be incorporated into the Shewhart-CUSUM control chart. Following the same restrictions as for prediction limits, they may be applied either to interwell or intrawell testing.

A disadvantage in applying control charts to groundwater monitoring data is that less is understood about their statistical performance, *i.e.*, false positive rates and power. The control limit used to identify potential releases to groundwater is not based on a formula incorporating a desired false positive rate (α). Unlike prediction limits, the control limit cannot be precisely set to meet a pre-specified SWFPR, unless the behavior of the control chart is modeled via Monte Carlo simulation. The same is true for assessing statistical power. Control charts usually provide less flexibility than prediction limits in designing a statistical monitoring program for a network.

In addition, Shewhart-CUSUM control charts are a parametric procedure with no existing non-parametric counterpart. Non-parametric prediction limit tests are still generally needed when the background data on which the control chart is constructed cannot be normalized. Control charts are mostly appropriate for analytes with a reasonably high detection frequency in monitoring wells. These include inorganic constituents (*e.g.*, detectable trace elements and geochemical monitoring parameters) occurring naturally in groundwater, and other persistently-found, site-specific chemicals.

6.5 SITE DESIGN EXAMPLES

Three hypothetical design examples consider a small, medium and large facility, illustrating the principles discussed in this chapter. In each example, the goal is to determine what statistical method or methods should be chosen and how those methods can be implemented in light of the two fundamental design criteria. Further design details are covered in respective **Part III** detection monitoring test

chapters, although very detailed site design is beyond the scope of the guidance. More detailed evaluations and examples of diagnostic tests are found in **Part II** of the guidance.

►EXAMPLE 6-5 SMALL FACILITY

A municipal landfill has 3 upgradient wells and 8 downgradient wells. Semi-annual statistical evaluations are required for five inorganic constituents. So far, six observations have been collected at each well. Exploratory analysis has shown that the concentration measurements appear to be approximately normal in distribution. However, each of the five monitored parameters exhibits significant levels of natural spatial variation from well to well. What statistical approach should be recommended at this landfill?

SOLUTION

Since the inorganic monitoring parameters are measurable and have significant spatial variability, it is recommended that parametric intrawell rather than interwell tests should be considered. Assuming that none of the downgradient wells is recently contaminated, each well has $n = 6$ observations available for its respective intrawell background. Six background measurements may or may not be enough for a sufficiently powerful test.

To address the potential problem of inadequate power, a one-way ANOVA should be run on the combined set of wells (including background locations). If the well-to-well variances are significantly different, individual standard deviation estimates should be made from the six observations at the eight downgradient wells. If the variances are approximately equal, a pooled standard deviation estimate can instead be computed from the ANOVA table. With 11 total wells and 6 measurements per well, the pooled standard deviation has $df = 11 \times 5 = 55$ degrees of freedom, instead of $df = 5$ for each individual well.

Regardless of ANOVA results, the per-test false positive rate is approximately the design SWFPR divided by the annual number of tests. For $w = 8$ compliance wells, $c = 5$ parameters monitored, and $n_E = 2$ statistical evaluations per year, the per-test false positive rate is approximately $\alpha_{test} = SWFPR/(w \times c \times n_E) = 0.00125$. Given normal distribution data, several different parametric prediction limit retesting plans can be examined,[17] using either the combined sample size of $df + 1 = 56$ or the per-well sample size of $n = 6$.

Explained in greater detail in **Chapter 19**, κ-multiples and power ratings for each test type (using the inputs $w = 8$ and $n = 6$ or 56 are obtained from the nine parametric **Appendix D** Intrawell tables labeled *'5 COC, Semi-Annual'*. The following κ-factors were obtained for tests of future values at $n = 6$: $\kappa = 3.46$ (1-of-2 test); $\kappa =$ **2.41** (1-of-3); $\kappa =$ **1.81** (1-of-4); and $\kappa = 2.97$ (modified California) plans. For future means, the corresponding κ-factors were: $\kappa = 4.46$ (1-of-1 mean size 2); $\kappa = 2.78$ (1-of-2 mean size 2); $\kappa = 2.06$ (1-of-3 mean size 2); $\kappa = 3.85$ (1-of-1 mean size 3); and $\kappa = 2.51$ (1-of-2 mean size 3). In these tables, *κ*-factors reported in **Bold** have good power, those *Italicized* have acceptable power and Plain Text indicates low power. For single well intrawell tests, only 1-of-3 or 1-of-4 plans for future values, 1-of-2 or 1-of-3 mean size 2 or 1-of-2 mean size 3 plans meet the ERPC criteria.

[17] Intrawell control charts with retesting are also an option, though the control limits associated with each retesting scheme need to be simulated.

Although each of these retesting plans is adequately powerful, a final choice would be made by balancing 1) the cost of sampling and chemical analysis at the site; 2) the ability to collect statistically independent samples should the sampling frequency be increased; and 3) a comparison of the actual power curves of the three plans. The last can be used to assess how differences in power might impact the rapid identification of a groundwater release. Since a 1-of-3 test for future observations has good power, it is unnecessary to make use of a 1-of-4 test. Similarly, the 1-of-3 test for mean size 2 and a 1-of-2 test for mean size 3 might also be eliminated, since a 1-of-2 test of a mean size 2 is more than adequate. This leaves the 1-of-3 future values and 1-of-2 mean 2 tests as the final prediction limit options to consider.

Though prediction limits around future means are more powerful than plans for observations, only 3 independent measurements might be required for a 1-of-3 test, while 4 might be necessary for the 1-of-2 test for mean size 2. For most tests at background, a single sample might suffice for the 1-of-3 test and 2 independent samples for the test using a 1-of-2 mean size 2.

Much greater flexibility is afforded if the pooled intrawell standard deviation estimate can be used. For this example, any of the nine parametric intrawell retesting plans is sufficiently powerful, including a 1-of-2 prediction limit test on observations and a 1-of-1 test of mean size 2. In order to make this assessment using the pooled-variance approach, a careful reading of **Chapter 13, Section 13.3**. is necessary to generate comparative κ-factors.

Less overall sampling is needed with the 1-of-2 plan on observations, since only a single sample may be needed for most background conditions. Two observations are always required for the 1-of-1 mean size 2 test. More prediction limit testing options are generally available for a small facility. ◄

►EXAMPLE 6-6 MEDIUM FACILITY

A medium-sized hazardous waste facility has 4 upgradient background wells and 20 downgradient compliance wells. Ten initial measurements have been collected at each upgradient well and 8 at downgradient wells. The permitted monitoring list includes 10 inorganic parameters and 30 VOCs. No VOCs have yet been detected in groundwater. The remaining 10 inorganic constituents are normal or can be normalized, and five show evidence of significant spatial variation across the site. Assume that pooled-variances cannot be obtained from the historical upgradient or downgradient well data. If one statistical evaluation must be conducted each year, what statistical method and approach are recommended?

SOLUTION

At this site, there are potentially 800 distinct well-constituent pairs that might be tested. But since none of the VOCs has been detected in groundwater in background wells, all 30 of the VOCs should be handled using the *double quantification rule* (**Section 6.2.2).** A second confirmatory resample should be analyzed at those compliance wells for any of the 30 VOC constituents initially detected. Two successive quantified detections above the RL are considered significant evidence of groundwater contamination at that well and VOC constituent. To properly limit the SWFPR, the 30 VOC constituents are excluded from further SWFPR calculations, which is now based on $w \times c \times n_E = 20 \times 10 \times 1 = 200$ annual tests.

The five inorganic constituent background data sets indicate insignificant spatial variation and can be normalized. The observations from the four upgradient wells can be pooled to form background data sets with an n = 40 for each of these five constituents. Future samples from the 20 compliance wells are then compared against the respective *interwell* background data. With one annual evaluation, c = 10 constituents, w = 20 wells and n = 40 background samples, the Interwell *'10 COC, Annual'* tables for parametric prediction limits with retesting can be searched in **Appendix D**. Alternatively, control chart limits can be fit to this configuration via Monte Carlo simulations. Even though only five constituents will be tested this way, all of the legitimate constituents (c) affecting the SWFPR calculation, are used in applying the tables.

Most of the interwell prediction limit retesting plans, whether for observations or means, offer good power relative to the annual evaluation ERPC. The final choice of a plan may be resolved by a consideration of sampling effort and cost, as well as perhaps a more detailed power comparison using simulated curves. For prediction limits, a 1-of-2 test for observations (κ = **2.18**) and the 1-of-1 prediction limit for a mean of order 2 (κ = **2.56**) both offer good power. These two plans also require the least amount of sampling to identify a potential release (as discussed in Example 6-6). Beyond this rationale, the more powerful 1-of-1 test of a future mean size 2 might be selected. Full power curves could be constructed and overlaid for several competing plans.

The remaining 5 inorganic constituents must be managed using intrawell methods based on individual compliance well sizes of n = 8. For the same c, w, and n_E inputs as above, the Appendix D Intrawell '10 COC, Annual' tables should be used. Only four of the higher order prediction limit tests have acceptable or good power: 1-of-4 future values (κ = **1.84**); 1-of-2 mean size 2 (κ = *2.68*); 1-of-3 mean size 2 (κ = **2.00**); and 1-of-2 mean size 3 (κ = **2.39**) tests. The 1-of-2 mean size 2 has only acceptable power. The first two tests require the fewest samples under most background conditions and in total, with the 1-of-4 test having superior power. ◄

►EXAMPLE 6-7 LARGE FACILITY

A larger solid waste facility must conduct two statistical evaluations per year at two background wells and 30 compliance wells. Parameters on the monitoring list include five trace metals with a high percentage of non-detect measurements, and five other inorganic constituents. While the inorganic parameters are either normal or can be normalized, a significant degree of spatial variation is present from one well to the next. If 12 observations were collected from each background well, but only 4 quarterly measurements from each compliance well, what statistical approach is recommended?

SOLUTION

Because the two groups of constituents evidence distinctly different statistical characteristics, each needs to be separately considered. Since the trace metals have occasional detections or 'hits,' they cannot be excluded from the SWFPR computation. Because of their high non-detect rates, parametric prediction limits or control charts may not be appropriate or valid unless a non-detect adjustment such as Kaplan-Meier or robust regression on order statistics is used (**Chapter 15**). Assuming for this example that parametric tests cannot be applied, the trace metals should be analyzed using non-parametric prediction limits. The presence of frequent non-detects may substantially limit the potential degree of spatial variation, making an *interwell* non-parametric test potentially feasible. The Kruskal-Wallis non-parametric ANOVA (**Chapter 17**) could be used to test this assumption.

In this case, the number of background measurements is n = 24, and this value along with w = 30 compliance wells would be used to examine possible non-parametric retesting plans in the **Appendix D** tables for non-parametric prediction limits. As these tables offer achievable per-evaluation, per-constituent false positive rates for each configuration of compliance wells and background levels, the target α level must be determined. Given semi-annual evaluations, the per-evaluation false positive rate is approximately $\alpha_E = 0.10/n_E = 0.05$. Then, with 10 constituents altogether, the approximate per-constituent false positive rate for each trace metal becomes $\alpha_{const} = 0.05/10 = 0.005$.

Only one retesting plan meets the target false positive rate, a 1-of-4 non-parametric prediction limit using the maximum value in background as the comparison limit. This plan has 'acceptable' power relative to the ERPC. Other more powerful plans all have higher-than-targeted false positive rates.

For the remaining 5 inorganic constituents, the presence of significant spatial variation and the fact that the observations can be normalized, suggests the use of parametric intrawell prediction or control limits. As in the previous **Example 6-6**, interwell prediction limit tables in **Appendix D** are used by identifying κ multipliers and power ratings based on *all* 10 constituents subject to the SWFPR calculations. This is true even though these parametric options only pertain to 5 constituents. The total number of well-constituent pair tests per year is equal to $w \times c \times n_E = 30 \times 10 \times 2 = 600$ annual tests.

Assuming none of the observed spatial variation is due to already contaminated compliance wells, the number of measurements that can be used as intrawell background per well is small (n = 4). A quick scan of the intrawell prediction limit retesting plans in **Appendix D** *'10COC, Semi-Annual'* tables indicates that none of the plans offer even acceptable power for identifying a potential release. A one-way ANOVA should be run on the combined set of w = 30 compliance wells to determine if a pooled intrawell standard deviation estimate can be used.

If levels of variance across these wells are roughly the same, the pooled standard deviation will have $df = w(n-1) = 30 \times 3 = 90$ degrees of freedom, making each intrawell prediction or control limit much more powerful. Using the **R** script provided in **Appendix C** for intrawell prediction limits with a pooled standard deviation estimate (see **Section 13.3**), based on n = 4 and df = 90, all of the relevant intrawell prediction limits are sufficiently powerful compared to the semi-annual ERPC. With the exception of the 1-of-2 future values test at acceptable power, the other tests have good power. The final choice of retesting plan can be made by weighing the costs of required sampling versus perhaps a more detailed comparison of the full power curves. Plans with lower sampling requirements may be the most attractive. ◄

CHAPTER 7. STRATEGIES FOR COMPLIANCE/ASSESSMENT AND CORRECTIVE ACTION

This chapter covers the fundamental design principles for compliance/assessment and corrective action statistical monitoring programs. One important difference between these programs and detection monitoring is that a fixed external GWPS is often used in evaluating compliance. These GWPS can be an MCL, risk-based or background limit as well as a remedial action goal. Comparisons to a GWPS in compliance/assessment and corrective action are generally *one-sample* tests as opposed to the two- or multi-sample tests in detection monitoring. Depending on the program design, *two- or multiple-sample* detection monitoring strategies can be used with well constituents subject to background compliance/corrective action testing. While a general framework is presented in this chapter, specific test applications and strategies are presented in **Chapters 21 and 22** for fixed GWPS comparisons. **Sections 7.1** through **7.4** discuss comparisons to fixed GWPSs, while **Section 7.5** covers background GWPS testing (either as a fixed limit or based on a background statistic). Discussions of regulatory issues are generally limited to 40 CFR Part 264, although they also apply to corresponding sections of the 40 CFR Part 258 solid waste rules.

7.1 INTRODUCTION

The RCRA regulatory structure for compliance/assessment and corrective action monitoring is outlined in **Chapter 2**. In detection and compliance/assessment monitoring phases, a facility is presumed not to be 'out of compliance' until significant evidence of an impact or groundwater release can be identified. In corrective action monitoring, the presumption is reversed since contamination of the groundwater has already been identified and confirmed. The null hypothesis of onsite contamination is rejected only when there is significant evidence that the clean-up or remediation strategy has been successful.

Compliance/assessment monitoring is generally begun when statistically significant concentration exceedances above background have been confirmed for one or more detection monitoring constituents. *Corrective action* is undertaken when at least one exceedance of a *hazardous* constituent GWPS has been identified in compliance/assessment monitoring. The suite of constituents subject to compliance/assessment monitoring is determined from Part 264 Appendix IX or Part 258 Appendix II testing, along with prior hazardous constituent data evaluated under the detection monitoring program. Following a compliance monitoring statistical exceedance, only a few of these constituents may require

the change in hypothesis structure to corrective action monitoring. This formal corrective action testing will need to await completion of remedial activities, while continued monitoring can track progress in meeting standards.

The same general statistical method of *confidence interval testing against a fixed GWPS* is recommended in both compliance/assessment and corrective action programs. As discussed more fully below and in **Chapter 21**, confidence intervals provide a flexible and statistically accurate method to test how a parameter estimated from a single sample compares to a fixed numerical limit. Confidence intervals explicitly account for variation and uncertainty in the sample data used to construct them.

Most decisions about a statistical program under §264.98 detection monitoring are tailored to facility conditions, other than selecting a target site-wide cumulative false positive rate and a scheme for evaluating power. Statistical design details are likely to be site-specific, depending on the available data, observed distributions and the scope of the monitoring network. For compliance/assessment and corrective action testing under §264.99 and §264.100 or similar tests against fixed health-based or risk-based standards, the testing regimen is instead likely to be determined in advance by the regulatory agency. The Regional Administrator or State Director is charged with defining the nature of the tests, constituents to be tested, and the wells or compliance points to be evaluated. Specific decisions concerning false positive rates and power may also need to be defined at a regulatory program level.

The advantage of a consistent approach for compliance/assessment and corrective action monitoring tests is that it can be applied across all Regional or State facilities. Facility-specific input is still needed, including the observed distributions of key constituents and the selection of statistical power and false positive criteria for permits. Because of the asymmetric nature of the risks involved, regulatory agency and facility perspectives may differ on which statistical risks are most critical. Therefore, we recommend that the following issues be addressed for compliance/assessment and corrective action monitoring (both §264.99 and §264.100), as well as for other programs involving comparisons to fixed standards:

- What are the appropriate hypothesis testing structures for making comparisons to a fixed standard?
- What do fixed GWPS represent in statistical terms and which population parameter(s) should be tested against them?
- What is a desirable frequency of sampling and testing, which test(s), and for what constituents?
- What statistical power requirements should be included to ensure protection of health and the environment?
- What confidence level(s) should be selected to control false positive error rates, especially considering sites with multiple wells and/or constituents?

Decisions regarding these five questions are complex and interrelated, and have not been fully addressed by previous RCRA guidance or existing regulations. This chapter addresses each of these points for both §264.99 and §264.100 testing. By developing answers at a regulatory program level, the necessity of re-evaluating the same questions at each specific site may be avoided.

7.2 HYPOTHESIS TESTING STRUCTURES

Compliance testing under §264.99 specifically requires a determination that one or more well constituents exceeds a permit-specific GWPS. The correct statistical hypothesis during compliance/assessment monitoring is that groundwater concentrations are presumed *not* to exceed the fixed standard unless sampling data from one or more well constituents indicates otherwise. The null hypothesis, H_0, assumes that downgradient well concentration levels are less than or equal to a standard, while the alternative hypothesis, H_A, is accepted only if the standard is significantly exceeded. Formally, for some parameter (Θ) estimated from sample data and representing a standard G, the relevant hypotheses under §264.99 compliance monitoring are stated as:

$$H_0 : \Theta \le G \; vs. \; H_A : \Theta > G \qquad [7.1]$$

Once a positive determination has been made that at least one compliance well constituent exceeds the fixed standard (*i.e.*, GWPS), the facility is subject to corrective action requirements under §264.100. At this point, the regulations imply and statistical principles dictate that the hypothesis structure should be *reversed* (for those compliance wells and constituents indicating exceedances). Other compliance constituents (*i.e.*, those not exceeding their respective GWPSs) may continue to be tested using equation 7.1 hypotheses. It is then assumed that contamination equal to or in excess of the GWPS exists and is presumed to be the case unless demonstrated otherwise. A positive determination that groundwater concentrations are below the standard is necessary to demonstrate regulatory compliance for any wells and constituents under remediation. In statistical terms, the relevant hypotheses for §264.100 are:

$$H_0 : \Theta \ge G \; vs. \; H_A : \Theta < G \qquad [7.2]$$

The reasoning behind this approach is as follows. Background exceedances by one or more well constituents under §264.98 detection monitoring do not predetermine any particular relationship of these increased concentration levels to fixed limits used as GWPS. Standards for different constituents vary over orders of magnitude. The actual concentration level triggering a statistically significant increase above background can vary considerably and bear little or no relationship to risk-based standards. Use of the initial compliance monitoring hypothesis framework in [7.1] ensures positive evidence that at least one hazardous constituent is truly above a GWPS. Since corrective action can be expensive and difficult, this provides important assurance that site program monitoring decisions are made correctly.

This guidance recognizes that not all regulatory programs are constructed alike. Objectives and regulatory interpretations may differ as to the basic goals of compliance/assessment or corrective action monitoring. When large numbers of sites with available hazardous constituent data are being screened to determine their need for remediation (perhaps outside the formal RCRA regulatory framework), the assessment may be conducted with the ***explicit presumption*** that contamination exists onsite. Presumably, elevated hazardous constituent concentrations have already been detected at these facilities. For these assessments, the compliance/assessment statistical hypothesis framework follows that presented in Equation 7.2. Instead of a *lower* confidence limit as recommended below, the appropriate statistical approach involves an ***upper*** confidence limit, as is appropriate for corrective action.

Non-RCRA programs seeking to use methods presented in the Unified Guidance may also presume a different statistical hypothesis structure from that presented here. The primary goal is to ensure that the statistical approach matches the appropriate hypothesis framework. It is also allowable under RCRA regulations to define GWPS based on background data, discussed further in **Section 7.5**.

Whatever the population parameter (Θ) selected as representative of the GWPS, testing consists of a confidence interval derived from the compliance point data at some choice of significance level (α), and then compared to the standard *G*. The confidence intervals describe the probable distribution of the sample statistic, θ, employed to estimate the true parameter Θ. For testing under compliance/assessment monitoring, a lower confidence limit around the true parameter — LCL(Θ) — is utilized. If LCL(Θ) exceeds the standard, there is statistically significant evidence in favor of the alternative hypothesis, H_A: $\Theta > G$, that the compliance standard has been violated. If not, the confidence limit test is inconclusive and the null hypothesis accepted.

When the corrective action hypothesis of [7.2] is employed, an *upper* confidence limit UCL(Θ) is generated from the compliance point data and compared to the standard *G*. In this case, the UCL(Θ) should lie *below* the standard to accept the alternative hypothesis that concentration levels are in compliance, H_A: $\Theta < G$. If the UCL(Θ) is larger than the standard, the test is inconclusive. It should be recognized that once corrective action or remediation activities are initiated, there will be a considerable time during which the GWPS may still be exceeded. As provided in the RCRA regulations, it is at the conclusion of remediation activities that formal corrective action monitoring evaluation is appropriate. However, in the intervening period of remedial activity, well constituents can still be monitored and the relative efficacy of remediation measures tracked. The same corrective action statistical hypotheses can be assumed for the targeted constituents; techniques such as trend testing may be appropriate interim applications.

If the entire confidence interval (considering both the lower and upper confidence limits) lies below the fixed standard *G* in either a compliance/assessment or corrective action setting, there is statistically significant evidence that the true parameter or characteristic (*e.g.*, the mean) is less than the standard. The constituent concentrations at the well are considered to be in compliance. Conversely, if the confidence interval lies entirely above *G*, the evidence suggests that the true parameter or characteristic exceeds the standard, and that concentrations at the well are out of compliance.

When the confidence interval straddles the standard *G* (as with the example confidence interval around the upper 95th percentile in **Figure 7-1** below), the correct decision is uncertain. When the population mean is being tested, and a confidence interval around the mean has accurately estimated its location, the true mean lies somewhere between the lower and upper confidence limits. But the *precise* value of the population mean within that range is unknown. The mean might be less than *G* or it might be greater than *G*. No clear decision with high statistical confidence is possible. ***When testing the compliance/assessment monitoring hypothesis of* [7.1]*, we recommend that the null hypothesis should not be rejected unless the entire confidence interval defined by and including the lower confidence limit exceeds the GWPS. By the same token, when testing the corrective action hypothesis of equation* [7.2]*, we recommend that the null hypothesis not be rejected unless the entire upper confidence interval and limit lies below the GWPS***.

These ideas can be illustrated with a normal confidence interval around the arithmetic mean. In this case, the population parameter Θ equals μ, the true population mean of a given compliance well

constituent. The statistic used to estimate μ is the sample mean ($\bar{x}$). With this statistic and normally-distributed data, the lower and upper confidence limits are symmetric:

$$LCL(\mu) = \bar{x} - t_{1-\alpha,n-1}\frac{s}{\sqrt{n}} \quad [7.3]$$

$$UCL(\mu) = \bar{x} + t_{1-\alpha,n-1}\frac{s}{\sqrt{n}} \quad [7.4]$$

for a selected significance level (α) and sample size *n*. Note in these formulas that *s* is the sample standard deviation, and $t_{1-\alpha,n-1}$ is a central Student's *t*-value with *n*–1 degrees of freedom.

The two hypothesis structures and tests are defined as follows:

Case A. Test of non-compliance (§264.99) *vs*. a fixed standard (compliance/assessment monitoring):

Test Hypothesis: $H_0 : \mu \le G \; vs. \; H_A : \mu > G$

Test Statistic: $LCL_{1-\alpha} = \bar{x} - t_{1-\alpha,n-1}\frac{s}{\sqrt{n}}$

Rejection Region: Reject null hypothesis (H_0) if $LCL_{1-\alpha} > G$; otherwise, accept null hypothesis

Case B. Test of compliance (§264.100) *vs*. a fixed standard (corrective action):

Test Hypothesis: $H_0 : \mu \ge G \; vs. \; H_A : \mu < G$

Test Statistic: $UCL_{1-\alpha} = \bar{x} + t_{1-\alpha,n-1}\frac{s}{\sqrt{n}}$

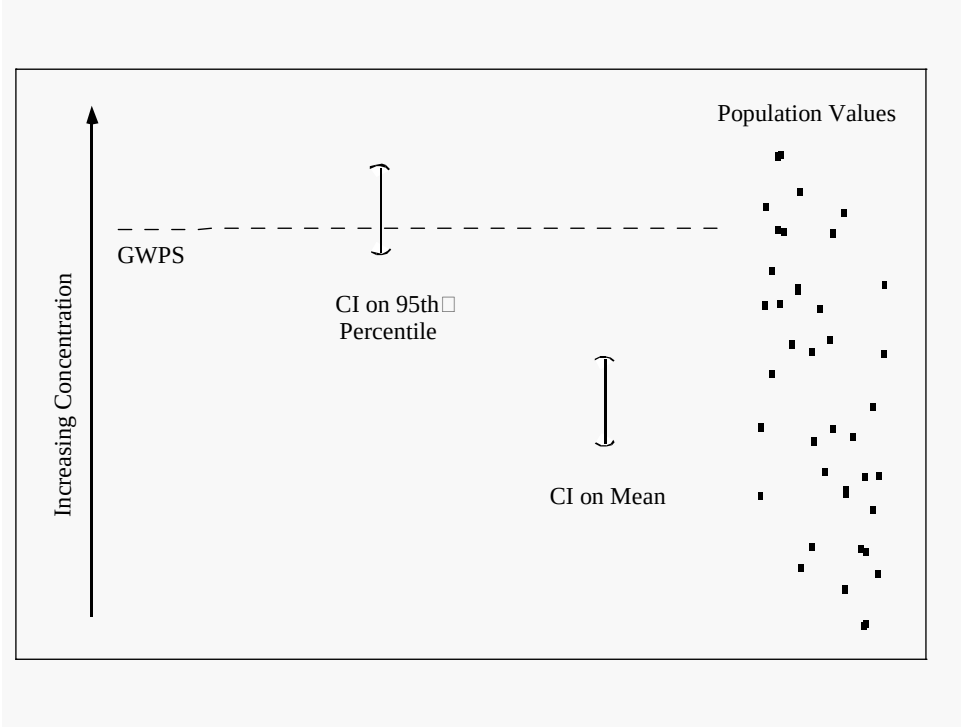
Rejection Region: Reject null hypothesis (H_0) if $UCL_{1-\alpha} < G$; otherwise, accept null hypothesis

For all confidence intervals and tests presented in **Chapters 21** and **22**, the test structures are similar to those above. But not every pair of lower and upper confidence limits (*i.e.*, LCL and UCL) will be symmetric, particularly for skewed distributions and in non-parametric tests on upper percentiles. For a non-parametric technique such as a confidence interval around the median, exact confidence levels will depend on the available sample size and which *order statistics* are used to estimate the desired population parameter. In these cases, an exact target confidence level may or may not be attainable.

When calculating confidence intervals, assignment of the false positive error (α) differs between a one-sided and two-sided confidence interval test. The symmetric upper and lower confidence intervals are shown in **Figure 7-1** largely for illustration purposes. If the *lower* confidence interval for some tested parameter Θ is the critical limit, all of the α error is assigned to the region below the LCL(Θ). Hence, a 1-α confidence level covers the range from the lower limit to positive infinity. Similarly, all of the α error for an upper confidence limit UCL(Θ) is assigned to the region above this value. For a two-

sided interval, the error rate is equally partitioned on both sides of the respective confidence interval limits. A 95% lower confidence limit implies that a 5% chance of an error exists for values lying below the limit. In contrast, a two-sided 95% confidence interval implies a 2.5% chance above and a 2.5% chance of an error below the confidence level. Depending on how confidence intervals are defined, the appropriate statistical adjustment (e.g., the *t*-value in **Equations 7-3** and **7-4**) needs to take this into account.

Figure 7-1. Confidence Interval on Mean vs. Fixed Upper Percentile Limit



7.3 GROUNDWATER PROTECTION STANDARDS

A second essential design step is to identify the appropriate population parameter and its associated statistical estimate. This is primarily a determination of what a given fixed GWPS represents in statistical terms. Not all fixed concentration standards are meant to represent the same statistical quantities. A distinction is drawn between 1) those central tendency standards designed to represent a mean or average concentration level and 2) those which represent either an upper percentile or the maximum of the concentration distribution. If the fixed standard represents an average concentration, it is assumed in the Unified Guidance that the ***mean*** concentration (or possibly the ***median*** concentration) in groundwater should not exceed the limit. When a fixed standard represents an ***upper percentile*** or ***maximum***, no more than a small, specified fraction of the individual concentration measurements should exceed the limit.

The choice of confidence interval should be based on the type of fixed standard to which the groundwater data will be compared. A fixed limit best representing an upper percentile concentration (***e.g.***, the upper 95th percentile) should not be compared to a confidence interval constructed around the arithmetic mean. Such an interval only estimates the location of the population mean, but says nothing about the specific upper percentile of the concentration distribution. The ***average*** concentration level

could be substantially less than the standard even though a significant fraction of the individual measurements exceeds the standard (see **Figure 7-1**).

There are a variety of fixed standards to which different statistical measures apply. Alternative GWPSs based on Agency risk-assessment protocols are cited as an option in the solid waste regulations at §258.55(i)(1). Many of the risk-assessment procedures identified in the CERCLA program make use of chronic, long-term exposure models for ingestion or inhalation. These procedures are identified in the (EPA, 1989b) Risk Assessment Guidance for Superfund (RAGS) and the Supplemental Guidance for Calculating the Concentration Term (EPA, 1992c), and serve as guidance for other EPA programs. In the latter document, the *arithmetic mean* is identified as the appropriate parameter for identifying environmental exposure levels. The levels are intended to identify chronic, time-weighted *concentration averages* based on lifetime exposure scenarios.

The primary *maximum contaminant levels* [MCL] promulgated under the Safe Drinking Water Act (SDWA) follow the same exposure evaluation principles. An MCL is typically based on 70-year risk-exposure scenarios (for carcinogenic compounds), assuming an ingestion rate of 2 liters of water per day at the average concentration over time. Similarly, long-term risk periods (*e.g.*, 6-years) are used for non-carcinogenic constituents, assuming average exposure concentrations. The promulgated levels also contain a safety multiplicative factor and are applied at the end-user tap. Calculations for ingestion exposure risk to soil contaminants by an individual randomly traversing a contaminated site are based on the average estimated soil concentration. It is expected that an exposed individual drinking the water or ingesting the soil is not afforded any protection in the form of prior treatment.

Other standards which may represent a population mean include some RCRA site permits that include comparisons against an *alternate concentration limit* [ACL] based on the average value of background data. In addition, some standards represent time-weighted averages used for carcinogenic risk assessments such as the *lifetime average daily dose* [LADD].

Fixed limits based explicitly on the *median concentration* include *fish ingestion exposure factors*, used in testing fish tissue for certain contaminants. The exposure factors represent the allowable concentration level below which at least half of the fish sample concentrations should lie, the 50th percentile of the observed concentration distribution. If this distribution is symmetric, the mean and median will be identical. For positively skewed populations, the mean concentration could exceed the exposure factor even though the median (and hence, a majority of the individual concentrations) is below the limit. It would therefore not be appropriate to compare such exposure factors against a confidence interval around the mean contaminant level, unless one could be certain the distribution was symmetric.

Fixed standards are sometimes based on *upper percentiles*. Scenarios of this type include risk-based standards designed to limit acute effects that result from short-term exposures to certain chemicals (*e.g.*, chlorine gas leaking from a rail car or tanker). There is greater interest in possible acute effects or transient exposures having a significant short-term risk. Such exposure events may not happen often, but can be important to track for monitoring and/or compliance purposes.

When even short exposures can result in deleterious health or environmental effects, the fixed limit can be specified as a maximum allowable concentration. From a statistical standpoint, the standard identifies a level which can only be exceeded a small fraction of the time (*e.g.*, the upper 90th percentile). If a larger than allowable fraction of the individual exposures exceeds the standard, action is

likely warranted, even if the average concentration level is below the standard. Certain MCLs are interpreted in this same manner; the term 'maximum' in maximum contaminant level would be treated statistically as an upper percentile limit. Examples include criteria for bacterial counts and nitrate/nitrite concentrations, best regarded as upper percentile limits.

As an example, exposure of infants to nitrate concentrations in excess of 10 mg/L (NO_3^- as N) in drinking water is a case where greater concern surrounds acute effects resulting from short-term exposure. The flora in the intestinal tract of infant humans and animals does not fully develop until the age of about six months. This results in a lower acidity in the intestinal tract, which permits the growth of nitrate reducing bacteria. These bacteria convert nitrate to nitrite. When absorbed into the bloodstream, nitrite interferes with the absorption of oxygen. Suffocation by oxygen starvation in this manner produces a bluish skin discoloration — a condition known as "blue baby" syndrome (or methemoglobinemia) — which can result in serious health problems, even death. In such a scenario, suppose that acute effects resulting from short-term exposure above some critical level should normally occur in no more than 10 percent of all exposure events. Then the critical level so identified would be equivalent to the upper 90th percentile of all exposure events.

Another example is the so-called 20-year flood recurrence interval for structural design. Flood walls and drainage culverts are designed to handle not just the average flood level, but also flood levels that have a 1 in 20 chance of being equaled or exceeded in any single year. A 20-year flood recurrence level is essentially equivalent to estimating the upper 95th percentile of the distribution of flood levels (***e.g.***, a flood of this magnitude is expected to occur only 5 times every 100 years).

The various limits identified as potential GWPS in **Chapter 2** pose some interpretation problems. §264.94 Table 1 values are identified as "Maximum Concentration[s] of Constituents for Groundwater Protection" for 14 hazardous constituents, originating from earlier Federal Water Pollution Control Administration efforts. While not a definitive protocol for comparison, it was indicated that the limits were intended to represent a concentration level that should not be exceeded most of the time. In an early Water Quality Criteria report (USDI, 1968), the language is as follows:

> "It is clearly not possible to apply these (drinking water) criteria solely as maximum single sample values. The criteria should *not be exceeded over substantial portions* of time."

Similarly, the more current MCLs promulgated under the SDWA are identified as "maximum contaminant limits". Even if the limits were derived from chronic, risk-based assessments, the same implication is that these limits should not be exceeded.

Individual EPA programs make sample data comparisons to MCLs using different approaches. For small-facility systems monitored under the SDWA, only one or two samples a year might be collected for comparison. Anything other than direct comparisons isn't possible. Some Clean Water Act programs use arithmetic comparisons (means or medians) rather than a fully statistical approach. CERCLA typically utilizes these standards in mean statistical comparisons, consistent with other chronic health-based levels derived from their program risk assessment equations. In short, EPA nationwide does not have a single operational definition or measure for assessing MCLs with sample data.

The Unified Guidance cannot directly resolve these issues. Since the regulations promulgated under RCRA presume the use of fully statistical measures for groundwater monitoring program

evaluations, the guidance provides a number of options for both centrality-based and upper limit tests. It falls upon State or Regional programs to determine which is the most appropriate parameter for comparison to a GWPS. As indicated above, the guidance does recommend that any operational definition of the appropriate parameter of comparison to GWPS's be applied uniformly across a program.

If a mean- or median-based centrality parameter is chosen, the guidance offers fairly straightforward confidence interval testing options. For a parameter representing some infrequent level of exceedance to address the "maximum" or "most" criteria, the program would need to identify a specific upper proportion and confidence level that the GWPS represents. Perhaps a proportion of 80 to 95% would be appropriate, at 90-95% confidence. It is presumed that the same standard would apply to both compliance and corrective action testing under §264.99 and §264.100. If non-parametric upper proportion tests must be used for certain data, very high proportions make for especially difficult tests to determine a return to compliance (**Chapter 22**) because of the number of samples required.

7.4 DESIGNING A STATISTICAL PROGRAM

7.4.1 FALSE POSITIVES AND STATISTICAL POWER IN COMPLIANCE/ASSESSMENT

As discussed in **Chapters 3 and 6**, the twin criteria in designing an acceptable detection monitoring statistical program are the site-wide false positive rate [SWFPR] and the effective power of the testing regimen. Both statistical measures are crucial to good statistical design, although from a regulatory perspective, ensuring adequate power to detect contaminated groundwater is of primary importance.

In compliance/assessment monitoring, statistical power is also of prime concern to EPA. There should be a high probability that the statistical test will positively identify concentrations that have exceeded a fixed, regulatory standard. In typical applications where a confidence interval is compared against a fixed standard, a low false positive error rate (α) is chosen without respect to the power of the test. Partly this is due to a natural desire to have high *statistical confidence* in the test, where ($1-\alpha$) designates the confidence level of the interval. But statistical confidence is *not* the same as power. The confidence level merely indicates how often — in repeated applications — the interval will contain the true population parameter (Θ); not how often the test will indicate an exceedance of a fixed standard. It has historically been much easier to select a single value for the false positive rate (α) than to measure power, especially since power is not a single number but a *function* of the level of contamination (as discussed in **Section 3.5**).

The power to detect increases above a fixed standard using a lower confidence limit can be negligible when contaminant variability is high, the sample size is small and especially when a high degree of confidence has been selected. To remedy this problem, the Unified Guidance recommends reversing the usual sequence: first select a desired level of power for the test ($1-\beta$), and then compute the associated (maximum) false positive rate (α). In this way, a pre-specified power can be maintained even if the sample size is too low to simultaneously minimize the risks of both Type I and Type II errors (*i.e.*, false positives and false negatives).

Specific methods for choosing power and computing false positive rates with confidence interval tests are presented in **Chapter 22**. Detailed applications of confidence interval tests are provided in **Chapter 21**. The focus here is on setting a basic framework and consistent strategies.

As noted above, selecting false positive error rates in compliance or assessment testing (§264.99) has traditionally been accomplished under RCRA by choosing a fixed, individual test α. This strategy is attractive if only for the sake of simplicity. Individual test-wise false positive rates in the range of α = .01 to α = .10 are traditional and easily understood. In addition, the Part 264 regulations in §264.97(i)(2) require a minimum individual false positive rate of α = .01 in both compliance and corrective action testing against a fixed standard, as well as in those tests not specifically exempted under detection monitoring.[1]

Given a *fixed* sample size and constant level of variation, the statistical power of a test method drops as the false positive rate decreases. A low false positive rate is often associated with low power. Since statistical power is of particular concern to EPA in compliance/assessment monitoring, somewhat higher false positive rates than the minimum α = .01 RCRA requirement may be necessary to maintain a pre-specified power over the range of sample sizes and variability likely to be encountered in RCRA testing situations. The key is sample variability. When the true population coefficient of variation [CV] is no greater than 0.5 (whether the underlying distribution is normal or lognormal), almost all lower confidence limit tests exhibit adequate power. When the variation is higher, the risk of false negative error is typically much greater (and thus the power is lower), which may necessitate setting a larger than usual individual α.

Based on the discussion regarding false positives in detection monitoring in **Chapter 6**, some might be concerned about the use of relatively high individual test-wise false positive rates (α) in order to meet a pre-specified power, especially when considering the cumulative false positive error rate across multiple wells and/or constituents (*i.e.*, SWFPR). Given that a number of compliance wells and constituents might need to be tested, the likelihood of occurrence of at least one false positive error increases dramatically. However, several factors specific to compliance/assessment monitoring need to be considered. Unlike detection monitoring where the number of tests is easily identified, the issue is less obvious for compliance/assessment or corrective action testing. The RCRA regulations do not clearly specify which wells and constituents must be compared to the GWPS in compliance/assessment monitoring other than wells at the 'compliance point.' In some situations, this has been interpreted to mean all compliance wells; in other instances, only at those wells with a documented exceedance.

While all hazardous constituents including additional ones detected in Part 264 Appendix IX monitoring are potentially subject to testing, many may still be at concentration levels insignificantly different from onsite background. Constituents without health-based limits may or may not be included in compliance testing. The latter would be tested against background levels, using perhaps an ACL computed as a *tolerance limit on background* (see **Section 7.5**). This also tends to complicate derivation of SWFPRs in compliance testing. It was also noted in **Section 7.2** that the levels at which contaminants are released bear no necessary relationship to fixed, health-based standards. In a typical release, some constituent levels from a suite of analytical parameters may lie orders of magnitude below their GWPS, while certain carcinogenic compounds may easily exceed their standards.

[1] In some instances, a test with "reasonable confidence" (that is, having adequate statistical power) for identifying compliance violations can be designed even if α < 0.01. This is particularly the case when the sample coefficient of variation is quite low, indicating small degrees of sample variability.

The simple example below illustrates typical low-level aquifer concentrations following a release of four common petrochemical facility hazardous organic constituents often detected together:

Analyte	Aquifer Concentration (ug/l)		MCL (ug/l)
	Mean	SD	
Benzene	20	10	5
Toluene	35	15	1,000
Ethylbenzene	40	20	700
Xylene	100	35	10,000

While benzene as a carcinogen has a very low health standard, the remaining three constituents have aquifer concentrations orders of magnitude lower than their respective MCLs. Realistically, only benzene is likely to impact the cumulative false positive rate in LCL testing. Similar relationships occur in releases measured by trace element and semi-volatile organic suites.

Even though the null hypotheses in detection and compliance/assessment monitoring are similar (and compound) in nature (see [7.1]), it is reasonable to presume in detection monitoring that the compliance wells have average concentrations *no less* than mean background levels.[2] Since it is these background levels to which the compliance point data are compared in the absence of a release, the compound null hypothesis in detection monitoring (H_0: $\mu_C \leq \mu_{BG}$) can be reformulated practically as (H_0: $\mu_C = \mu_{BG}$). In this framework, individual concentration measurements are likely to occasionally exceed the background average and at times cause false positives to be identified even when there has been no change in average groundwater quality.

In compliance/assessment monitoring, the situation is generally different. The compound null hypothesis (H_0: $\mu_C \leq$ GWPS) will include some wells and constituents where the sample mean equals or nearly equals the GWPS when testing begins. But many well-constituent pairs may have true means considerably less than the standard, making false positives much less likely for those comparisons and lowering the overall SWFPR. How much so will depend on both the variability of each individual constituent and the degree to which the true mean (or relevant statistical parameter Θ) is lower than the GWPS for that analyte.

Because of this, determining the relevant number of comparisons with non-negligible false positive error rates may be quite difficult. The SWFPR in this situation would be defined as the probability that at least one or more lower confidence limits exceeded the fixed standard *G*, when the true parameter Θ (usually the mean) was actually below the standard. However, the relevant number of comparisons will depend on the nature and extent of the release. For a more extensive release, there is greater likelihood that the null hypothesis is no longer true at one or more wells. Instead of computing false positive rates, the focus should shift to minimizing false negative errors (*i.e.*, the risk of missing contamination above the GWPS).

[2] Note that background might consist of early intrawell measurements from compliance wells when substantial spatial variability exists.

On balance, the Unified Guidance considers computation of cumulative SWFPRs in compliance/assessment testing to be problematic, and reliance on individual test false positive rates preferable. The above arguments also suggest that flexibility in setting individual test-wise α levels may be appropriate.

7.4.2 FALSE POSITIVES AND STATISTICAL POWER IN CORRECTIVE ACTION

When contamination above a GWPS is confirmed, corrective action is triggered. Following a period of remediation activity, formal statistical testing will usually involve an *upper* confidence limit around the mean or an upper percentile compared against a GWPS. EPA's overriding concern in corrective action is that remediation efforts not be declared successful without sufficient statistical proof. Since groundwater is now presumed to be impacted at unacceptable levels, a facility should not exit corrective action until there is sufficient evidence that contamination has been abated.

Given the reversal of test hypotheses from compliance/assessment monitoring to corrective action (*i.e.*, comparing equation [7.1] with [7.2]), there is an asymmetry in regulatory considerations of false positive and false negative rates depending on the stage of monitoring. In compliance/assessment monitoring using tests of the lower confidence limit, the principal regulatory concern is that a given test has adequate statistical power to detect exceedances above the GWPS.

Permitted RCRA monitoring is likely to involve small annual well sample sizes based on quarterly or semi-annual sampling. To meet a pre-specified level of power by controlling the false negative rate (β) necessitates varying the false positive rate (α) for individual tests. Controlling an SWFPR for these tests (using a criterion like the SWFPR) is usually not practical because of the ambiguity in identifying the relevant number of potential tests and the difficulty of properly assigning via the subdivision principle (**Chapter 19**) individual fractions of a targeted SWFPR.

By contrast under corrective action using an *upper* confidence limit for testing, the principal regulatory and environmental concern is that one or more constituents might falsely be declared below a GWPS in concentration. Under the corrective action null hypothesis [7.2] this would be a *false positive error*, implying that α should be minimized during this sort of testing, instead of β. Specific methods for accomplishing this goal are presented in **Chapter 22**.

A remaining question is whether SWFPRs should be controlled during corrective action. While potentially desirable, the number of well-constituent pairs exceeding their respective GWPS and subject to corrective action testing is likely to be small relative to compliance testing. Not all compliance wells or constituents may have been impacted, and some may not be contaminated to levels exceeding the GWPS, depending on the nature, extent, and intensity of the plume. Remediation efforts would focus on those constituents exceeding their GWPS.

As noted in **Section 7.4.1**, the tenuous relationship between ambient background levels, contaminant magnitudes, and risk-based health standards implies that most GWPS exceedances are likely to be carcinogens, usually representing a small portion of all monitored constituents. Some exceedances may also be related compounds, for instance, chlorinated hydrocarbon daughter degradation products.

Statistically, the fact that some wells are contaminated while others may not be makes it difficult to define SWFPRs in corrective action. Instead, the Unified Guidance attempts to limit the individual test-wise α at those wells where exceedances have been confirmed and that are undergoing remediation. Since the most important consideration is to ensure that the true population parameter (Θ) is actually below the clean-up standard before declaring remediation a success, this guidance recommends the use of a reasonably low, *fixed* test-wise false positive rate (*e.g.*, α = .05 or .10). Under this framework, there will be a 5% to 10% chance of incorrectly declaring any single well-constituent pair of being in compliance when its concentrations are truly above the remedial standard.

The regulatory position in corrective action concerning statistical power is one of relative indifference. Although power under [7.2] represents the probability that the confidence interval test will correctly identify concentrations to be below the regulatory standard when in fact they are, the onus of proof for demonstrating that remediation has succeeded (*e.g.*, μ_C < GWPS) falls on the regulated facility. As it is the facility's interest to demonstrate compliance, it may wish to develop statistical power criteria which would enhance this possibility (including increasing test sample sizes).

7.4.3 RECOMMENDED STRATEGIES

As noted in **Section 7.1**, the Unified Guidance recommends the use of confidence intervals in both compliance/assessment and corrective action testing. In compliance/assessment, the lower confidence limit is the appropriate statistic of interest, while in corrective action it is the upper confidence limit. In either case, the confidence limit is compared against a fixed, regulatory standard as a one-sample test. These recommendations are consistent with good statistical practice, as well as literature in the field, such as Gibbons and Coleman (2001).

The type of confidence interval test will initially be determined by the choice of parameter(s) to represent the GWPS (**Section 7.2**). While this discussion has suggested that the mean may be the most appropriate parameter for chronic, health-based limits, other choices are possible. **Chapter 21** identifies potential test statistical tests of a mean, median or upper percentile as the most appropriate parameters for comparison to a GWPS. In turn, data characteristics will determine whether parametric or non-parametric test versions can be used. Depending on whether normality can be assumed for the original data or following transformation, somewhat different approaches may be needed. Finally, the presence of data trends affects how confidence interval testing can be applied.

Some regulatory programs prefer to compare each *individual* measurement against *G*, identifying a well as out-of-compliance if any of the individual concentrations exceeds the standard. However, the false positive rate associated with such strategies tends to be quite high if the parameter choice has not been clearly specified. Using this individual comparison approach and assuming a mean as the parameter of choice, is of particular concern. If the true mean is *less than but close to* the standard, chances are very high that one or more individual measurements will be greater than the limit even though the hypothesis in [7.1] has not been violated. Corrective action could then be initiated on a false premise. To evaluate whether a limited number of sample data exceed a standard, a lower confidence interval test would need to be based on a pre-specified upper percentile assumed to be the appropriate parameter for comparison to the GWPS.

Small individual well sample sizes and data uncertainty can rarely be avoided in compliance/assessment and corrective action. Given the nature of RCRA permits, sampling frequencies

in compliance/assessment or corrective action monitoring are likely to be established in advance. Relatively small sample sizes per well-constituent pair each year are likely to be the rule; the Unified Guidance assumes that quarterly and semi-annual sampling will be very typical.

For small and highly variable sample data sets, compliance/assessment monitoring and corrective action tests will have low statistical power either to detect exceedances above fixed standards or to demonstrate compliance in corrective action. One way to both enhance statistical power and control false positive error rates is through *incremental* or *sequential pooling* of compliance point data over time. Adding more data into a test of non-compliance or compliance will generally result in narrower confidence intervals and a clearer decision with respect to a compliance standard.

The Unified Guidance recommends accumulating compliance data over time at each well, by allowing construction of confidence limits on *overlapping* as opposed to *distinct* or *mutually exclusive* data sets. If the lower confidence limit [LCL] exceeds the GWPS in compliance/assessment, a clear exceedance can be identified. If the upper confidence limit [UCL] is below the GWPS in corrective action, remediation at that well can be declared a success. If neither of these respective events occurs, further sampling should continue. A confidence interval can be recomputed after each additional 1 or 2 measurements and a determination made whether the position of the confidence limit has changed relative to the compliance standard.

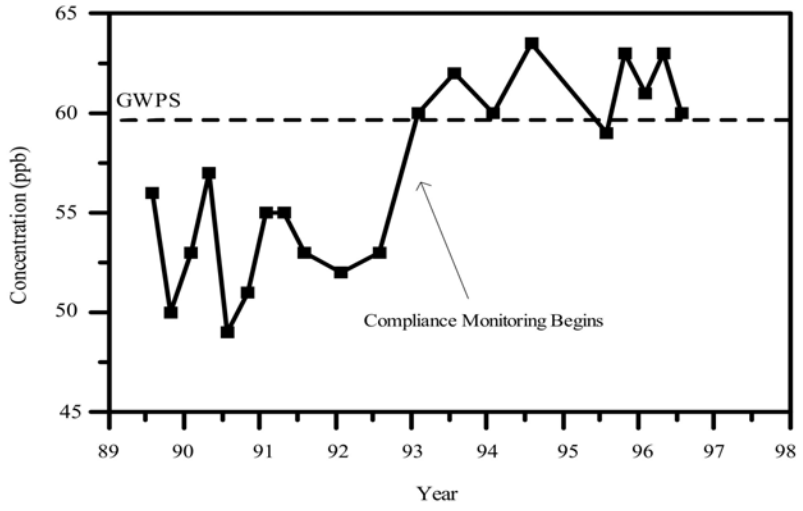
Tests constructed in this way at each successive evaluation period will not be statistically independent; instead, the proposed testing strategy falls into the realm of *sequential analysis*. But it should help to minimize the possibility that a small group of spurious values will either push a facility into needless corrective action or prevent a successful remedial effort from being identified.

One caveat with this approach is that it must be reasonable to assume that the population parameter Θ is stable over time. If a release has occurred and a contaminant plume is spreading through the aquifer, concentration shifts in the form of increasing trends over time may be more likely at contaminated wells. Likewise under active remediation, decreasing trends for a period of time may be more likely. Therefore, it is recommended that the sequential testing approach be used *after* aquifer conditions have stabilized to some degree. While concentration levels are actively changing with time, use of confidence intervals around a trend line should be pursued (see **Section 7.4.4** and **Chapter 21**).

7.4.4 ACCOUNTING FOR SHIFTS AND TRENDS

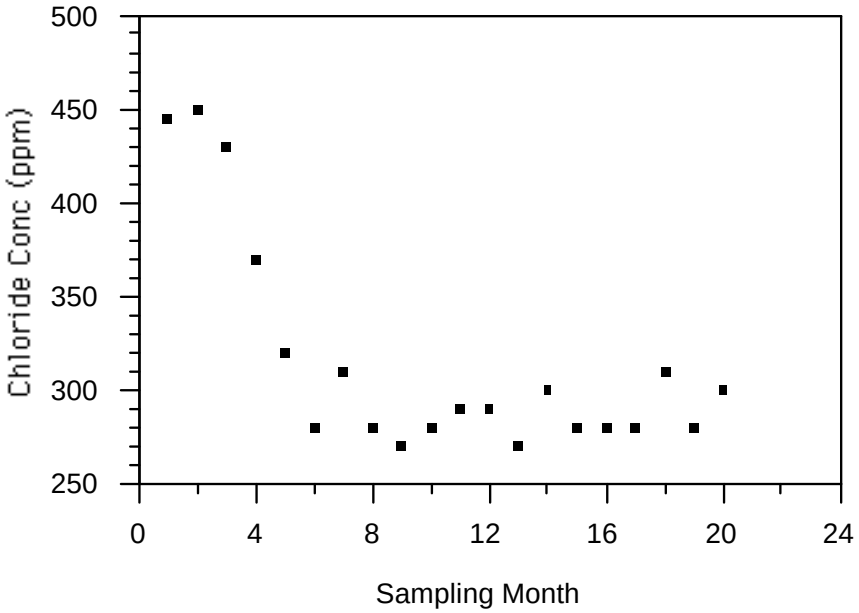
While accumulating compliance point data over time and successively re-computing confidence limits is appropriate for stable (*i.e.*, stationary) populations, it can give misleading or false results when the underlying population is changing. Should a release create an expanding contaminant plume within the aquifer, concentration levels at some or all of the compliance wells will tend to shift upward, either in discrete jumps (as illustrated in **Figure 7-2**) or an increasing trend over time. In these cases, a lower confidence limit constructed on accumulated data will be overly wide (due to high sample variability caused by combining pre- and post-shift data) and not be reflective of the more recent upward shift in the contaminant distribution.

Figure 7-2. Effect on Confidence Intervals of Stable Contamination Level



A similar problem can arise with corrective action data. Aquifer modifications as part of contaminant removals are likely to result in observable declines in constituent concentrations during the active treatment phase. At some point following cessation of remedial action, a new steady-state equilibrium may be established (**Figure 7-3**).

Figure 7-3. Decreasing Trend During Corrective Action

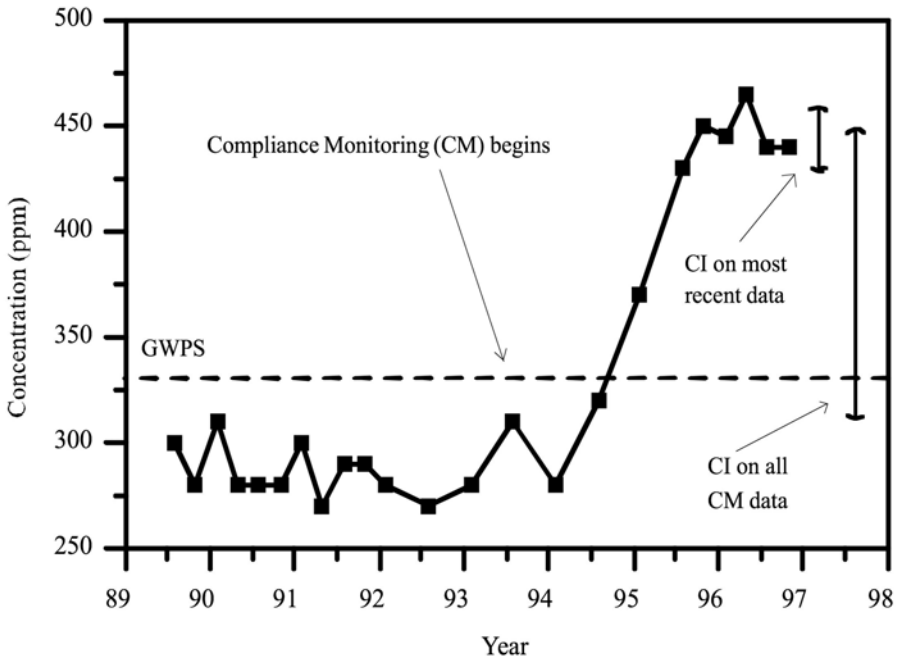


Until then, it is inappropriate to use a confidence interval test around the mean or an upper percentile to evaluate remedial success with respect to a clean-up standard. During active treatment phases and under non-steady state conditions, other forms of analysis such as confidence bands around a trend (see below), are recommended and should be pursued.

The Unified Guidance considers two basic types of non-stationary behavior: shifts and (linear) trends. A shift refers to a significant mean concentration increase or decrease departing from a roughly stable mean level. A trend refers to a series of consecutive measurements that evidence successively increasing or decreasing concentration levels. More complicated non-random data patterns are also possible, but beyond the scope of this guidance. With these two basic scenarios, the strategy for constructing an appropriate confidence interval differs.

An important preliminary step is to track the individual compliance point measurements on a time series plot (**Chapter 9**). If a discrete shift in concentration level is evident, a confidence limit should be computed on the most recent stable measurements. Limiting the observations in this fashion to a specific time period is often termed a 'moving window.' The reduction in sample size will often be more than offset by the gain in statistical power. More recent measurements may exhibit less variation around the shifted mean value, resulting in a shorter confidence interval (**Figure 7-4**). The sample size included in the moving window should be sufficient to achieve the desired statistical power (compliance/assessment) or false positive rate (corrective action). However, measurements that are clearly unrepresentative of the newly shifted distribution should not be included, even if the sample size suffers. Once a stable mean can be assumed, the strategy of sequential pooling can be used.

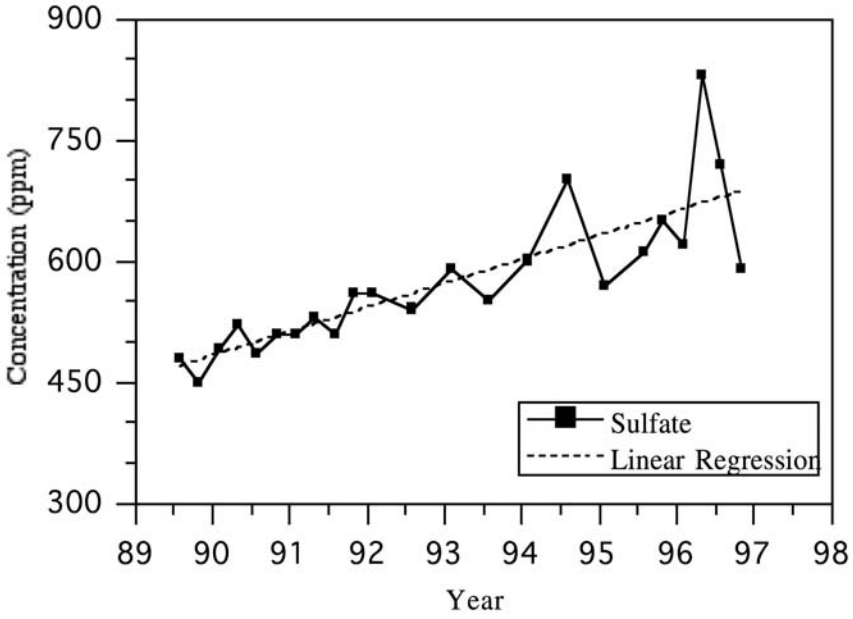
Figure 7-4. Effect on Confidence Intervals of Concentration Shift



If well concentration levels exhibit an increasing or decreasing trend over time (such as the example in **Figure 7-5**) and the pattern is reasonably linear or monotone, the trend can be identified using the methods detailed in **Chapter 17**. To measure compliance or non-compliance, a *confidence band* can be constructed around the estimated trend line, as described in **Chapter 21**. A confidence band is essentially a continuous series of confidence intervals estimated along every point of the trend. Using this technique, the appropriate upper or lower confidence limits at one or more points in the most recent

portion of the end of the sampling record can be compared against the fixed standard. The lower band is used to determine whether or not an exceedance has occurred in compliance/assessment, and an upper confidence band to determine if remedial success has been achieved in corrective action.

Figure 7-5. Rising Trend During Compliance Monitoring



By explicitly accounting for the trend, the confidence interval in **Chapter 21** will adjust upward or downward with the trend and thus more accurately estimate the current true concentration levels. Trend techniques are not just used to track progress towards exceeding or meeting a fixed standard. Confidence bands around the trend line can also provide an estimate of confidence in the average concentration as it changes over time. This subject is further covered in the Comprehensive Environmental Response, Compensation, and Liability Act [CERCLA] guidance ***Methods for Evaluating the Attainment of Cleanup Standards — Volume 2: Groundwater*** (EPA, 1992a).

A final determination of remedial success should not solely be a statistical decision. In many hydrologic settings, contaminant concentrations tend to rise after groundwater pumping wells are turned off due to changes in well drawdown patterns. Concentration levels may exhibit more complicated behavior than the two situations considered above. Thus, on balance, it is recommended that determining achievement of corrective action goals be done in consultation with the site manager, geologist, and/or remedial engineer.

7.4.5 IMPACT OF SAMPLE VARIABILITY, NON-DETECTS, AND NON-NORMAL DATA

Selection of hazardous constituents to be monitored in compliance/assessment or corrective action is largely determined by permit decisions. Regulatory requirements (*e.g.*, Part 264, Appendix IX) may also dictate the number of constituents. As a practical matter, the most reliable indicators of contamination should be favored. Occasionally, constituents subject to degradation and transformation in the aquifer (*e.g.*, chlorinated hydrocarbon suites) may result in additional, related constituents of concern.

Since health-based considerations are paramount in this type of monitoring, the most sensitive constituents from a health risk standpoint could be selected. But even with population parameters (Θ), sample sizes, and constituents determined, selecting an appropriate confidence interval test from **Chapter 21** can be problematic. For mildly variable sample data, measured at relatively stable levels, tests based on the normal distribution should be favored, whether constructed around a mean or an upper percentile. With highly variable sample data, selection of a test is less straightforward. If the observed data happen to be lognormal, Land's confidence interval around the arithmetic mean is a valid option; however, it has low power to measure compliance as the observations become more variable, and upward adjustment of the false positive rate (α) may be necessary to maintain sufficient power.

In addition, the extreme variability of an upper confidence limit using Land's technique can severely restrict its usage in tests of compliance during corrective action. Depending on the data pattern observed, degree of variability, and how closely the sample mimics the lognormal model, consultation with a professional statistician should be considered to resolve unusual cases. When the lognormal coefficient of variation is quite high, one alternative is to construct an upper confidence limit around the lognormal geometric mean (**Chapter 21**). Although such a confidence limit does not fully account for extreme concentration values in the right-hand tail of the lognormal distribution, a bound on the geometric mean will account for the bulk of possible measurements. Nonetheless, use of a geometric mean as a surrogate for the population arithmetic mean leads to distinctly different statistical test characteristics in terms of power and false positive rates.

In sum, excessive sample variability can severely limit the effectiveness of traditional compliance/assessment and corrective action testing. On the other hand, if excessive variability is primarily due to trends observable in the data, confidence bands around a linear trend can be constructed (**Section 7.4.4**).

LEFT-CENSORED SAMPLES

For compliance point data sets containing left-censored measurements (*i.e.*, non-detects), parametric confidence intervals cannot be computed directly without some adjustment. All of the parametric confidence intervals described in **Chapter 21** require estimates of the population mean μ and standard deviation σ. A number of adjustment strategies are presented in **Chapter 15.** If the percentage of non-detects is small — no more than 10-15% — *simple substitution* of half the reporting limit [RL] for each non-detect will generally work to give an approximately correct confidence interval.

For samples of at least 8-10 measurements and up to 50% non-detects, the *Kaplan-Meier* or *robust regression on order statistics* [ROS] methods can be used. Data should first be assessed via a *censored probability plot* whether the sample can be normalized. If so, these techniques can be used to compute estimates of the mean μ and standard deviation σ adjusted for the presence of left-censored values. These adjusted estimates can be used in place of the sample mean ($\bar{x}$) and standard deviation (s) listed in the confidence interval formulas of **Chapter 21** around either a mean or upper percentile.

If none of these adjustments is appropriate, non-parametric confidence intervals on either the median or an upper percentile (**Section 21.2**) can be calculated. Larger sample sizes are needed than with parametric confidence interval counterparts, especially for intervals around an upper percentile, to ensure a high level of confidence and a sufficiently narrow interval. The principal advantage of non-parametric

intervals is their flexibility. Not only can large fractions of non-detects be accommodated, but non-parametric confidence intervals can also be applied to data sets which cannot be normalized.

For heavily censored small data sets of 4-6 observations, the options are limited. One approach is to replace each non-detect by half its RL and compute the confidence interval *as if* the sample were normal. Though the resulting interval will be approximate, it can provide a preliminary indication of the well's compliance with the standard until further sampling data can be accumulated and the confidence interval recomputed.

Confidence bands around a trend can be constructed with censored data using a bootstrapped Theil-Sen non-parametric trend line (**Section 21.3.2**). In this method, the Theil-Sen trend is first computed using the sample data, accounting for the non-detects. Then a large number bootstrap resamples are drawn from the original sample, and an alternate Theil-Sen trend is conducted on each bootstrap sample. Variability in these alternate trend estimates is then used to construct a confidence band around the original trend.

LOGNORMAL AND OTHER NORMALIZED DATA

Lognormal data may require special treatment when building a confidence interval around the mean. Land's method (**Section 21.1.3**) can offer a reasonable way to accommodate the transformation bias associated with the logarithm, particularly when computing a lower confidence limit as recommended in compliance/assessment monitoring. For data normalized by transformations other than the logarithm, one option is to calculate a normal-based confidence interval around the mean using the transformed measurements, then back-transform the limits to the original concentration scale. The resulting interval will *not* represent a confidence interval around the arithmetic mean of the original data, but rather will estimate the confidence intervals of the *median* and/or *geometric mean*.

If the difference between the arithmetic mean and median is not considered important for a given GWPS, this strategy will be the easiest to implement. A wide range of results can occur with Land's method on highly skewed lognormal populations especially when computing an upper confidence limit around the arithmetic mean (Singh *et al.*, **1997**). It may be better to either construct a confidence interval around the lognormal *geometric* mean (**Section 21.1.2**) or to use the technique of *bootstrapping* (Efron, 1979; Davison and Hinkley, 1997) to create a non-parametric interval around the arithmetic mean.[3]

For confidence intervals around an upper percentile, no bias is induced by data that have been normalized via a transformation. Whatever the transformation used (*e.g.*, logarithm, square root, cube, *etc.*), a confidence interval can be constructed on the transformed data. The resulting limits can then be back-transformed to provide confidence limits around the desired upper percentile in the concentration domain.

[3] Bootstrapping is widely available in statistical software, including the open source **R** computing environment and EPA's free-of-charge **ProUCL** package. In some cases, setting up the procedure correctly may require professional statistical consultation.

7.5 COMPARISONS TO BACKGROUND DATA

Statistical tests in compliance/assessment and corrective action monitoring will often involve a comparison between compliance point measurements and a promulgated fixed health-based limit or a risk-based remedial action goal as the GWPS, described earlier. But a number of situations arise where a GWPS must be based on a background limit. The Part 264 regulations presume such a standard as one of the options under §264.94(a); an ACL may also be determined from background under §264.94(b). More recent Part 258 rules specify a background GWPS where a promulgated or risk-based standard is not available or if the historical background is greater than an MCL [§258.55(h)(2) & (3)].

Health-based risk standards bear no necessary relationship to site-specific aquifer concentration levels. At many sites this poses no problem, since the observed levels of many constituents may be considerably lower than their GWPS. However, either naturally-occurring or pre-existing aquifer concentrations of certain analytes can exceed promulgated standards. Two commonly monitored trace elements in particular-- arsenic and selenium-- are occasionally found at uncontaminated background well concentrations exceeding their respective MCLs. The regulations then provide that a GWPS based on background levels is appropriate.

A number of factors should be considered in designing a background-type GWPS testing program for compliance/assessment or corrective action monitoring. The most fundamental decision is whether to base such comparisons on *two- (or multiple-) sample* versus *single-sample* tests. For the first, many of the design factors discussed for detection monitoring in **Chapter 6** will be appropriate; for single sample comparisons to a fixed background GWPS, a confidence level approach similar to that discussed earlier for testing fixed health standards in this **Chapter 7** would be applied. This basic decision then determines how the GWPS is defined, the appropriate test hypotheses, types of statistical tests, what the background GWPS represents in statistical terms, and the relevance of individual test and cumulative false positive error rates. Such decisions may also be constrained by State groundwater anti-degradation policies. Other design factors to consider are the number of wells and constituents tested, interwell versus intrawell options, background sample sizes, and power. Unlike a single fixed standard like an MCL, background GWPS's may be uniquely defined for a given monitoring well constituent by a number of these factors.

SINGLE- VERSUS TWO-SAMPLE TESTING

One of two fundamental testing approaches can be used with site-specific background GWPSs. Either 1) a GWPS is defined as the critical limit from a pre-selected detection-level statistical test (e.g., a prediction limit) based on background measurements, or 2) background data are used to generate a fixed GWPS somewhat elevated above current background levels. In both cases, the resulting GWPS will be constituent- and possibly compliance well- specific. The first represents a *two-sample test* of two distinct populations (or more if a multiple-sample test) similar to those utilized in detection monitoring. As such, the individual test false positive rate, historical background sample size, cumulative false positive considerations, number of annual tests and desired future sample size will uniquely determine the limit. Whatever the critical value for a selected background test, it becomes the GWPS under compliance/assessment or corrective action monitoring.

The only allowable hypothesis test structure for the two-sample approach follows that of detection and compliance monitoring [7.1]. Once exceeded and in corrective action, a return to compliance is through evidence that future samples lie below the GWPS using the same hypothesis structure.

The second option uses a fixed statistic from the background data as the GWPS in a ***single-sample*** confidence interval test. Samples from a single population are compared to the fixed limit. In other respects, the strategy follows that outlined in **Chapter 7** for fixed health- or risk-based GWPS tests. The compliance/assessment test hypothesis structure also follows [7.1], but the hypotheses are reversed as in [7.2] for corrective action testing.

The choice of the single-sample GWPS deserves careful consideration. In the past, many such standards were simply computed as multiples of the background sample average (***i.e.***, GWPS = $2 \cdot \bar{x}$). However, this approach may not fully account for natural variation in background levels and lead to higher than expected false positive rates. If the GWPS were to be set at the historical background sample mean, even higher false positive rates would occur during compliance monitoring, and demonstrating corrective action compliance becomes almost impossible.

In the recommendations which follow below, an upper tolerance limit based on both background sample size and sample variability is recommended for identifying the background GWPS at a suitably high enough level above current background to allow for reversal of the test hypotheses. Although a somewhat arbitrary choice, a GWPS based on this method allows for a variety of confidence interval tests (e.g., a one-way normal mean confidence interval identified in equations [7.3] and [7.4]).

WHAT A BACKGROUND GWPS REPRESENTS

If the testing protocol involves two-sample comparisons, the background GWPS is an upper limit statistical interval derived from a given set of background data based on one or another detection monitoring tests discussed in **Chapter 6** and detailed in **Part III**. In these cases, the appropriate testing parameter is the true *mean* for the parametric tests, and the true *median* for non-parametric tests. This would include 1-of-*m* prediction limit detection tests involving future values. If a single-sample comparison against a fixed background GWPS is used, the appropriate parameter will also depend upon the type of confidence interval test to be used (**Part IV**). Except for parametric or non-parametric upper percentile comparisons, the likely statistical parameter would again be a mean (arithmetic, logarithmic, geometric) or the median. A background GWPS could be defined as an upper percentile parameter, making use of normal test confidence interval structures found in **Section 21.1.4**. Non-parametric percentile options would likely require test sample sizes too large for most applications. The Unified Guidance recommended approaches for defining single-sample GWPSs discussed later in this section presume a central tendency test parameter like the mean or median.

NUMBER OF MONITORED WELLS AND CONSTITUENTS

Compliance/assessment or corrective action monitoring tests against a fixed health- or risk-based standard (including single-sample background GWPSs) are not affected in a significant manner by the number of annual tests. But this would not be true for two- or multiple-sample background GWPS testing. In similar fashion to detection monitoring, the total number of tests is an important consideration in defining the appropriate false positive error test rate (α_{test}). The total number of annual tests is determined by how many compliance wells, constituents and evaluations occur per year.

Regulatory agency interpretations will determine the number and location of compliance monitoring wells. These can differ depending on whether the wells are unit-specific, and if a reasonable subset can be shown to be affected by a release. Perhaps only those compliance wells containing detectable levels of a compliance monitoring constituent need be included. Formal annual tests are generally required semi-annually, but other approaches may be applied.

The number of constituents subject to two-sample background GWPS testing will also depend on several factors. Only ***hazardous*** constituents not having a health- or risk-based standard are considered here. The basic criterion in interpreting required Part 264 Appendix IX or Part 258 Appendix II analyses is to identify those hazardous constituents found in downgradient compliance wells. Some initially detected common laboratory or sampling contaminants might be eliminated following a repeat scan. The remainder of the qualifying constituents will then require some form of background GWPS's. Along with the number of wells and annual evaluations, the total annual number of background tests will then be used in addressing an overall design cumulative design false positive rate.

In corrective action testing (for either the one- or two-sample approaches), the number of compliance wells and constituents may differ. Only those wells and constituents showing a significant compliance test exceedance might be used. However, from a standpoint of eventually demonstrating compliance under corrective action, it might be appropriate to still use the compliance/assessment GWPS for two-sample tests. With single-sample tests, the GWPS is compared individually by well and constituent as described.

BACKGROUND SAMPLE SIZES and INTERWELL vs. INTRAWELL TESTING

Some potential constituents may already have been monitored during the detection phase, and have a reasonable background size. Others identified under Part 264 Appendix IX or part 258 Appendix II testing may have no historical background data bases and require a period of background sampling.

Historical constituent well data patterns and the results of this testing may help determine if an interwell or intrawell approach should be used for a given constituent. For example, if arsenic and selenium were historical constituents in detection monitoring, they might also be identified as candidates for compliance background GWPS testing. There may already be indications that individual well spatial differences will need to be taken into account and an intrawell approach followed. In this case, individual compliance well background GWPSs need to be established and tested. On the other hand, certain hazardous trace elements and organics may only be detected and confirmed in one or more compliance wells with non-detects in background upgradient wells and possibly historical compliance well data. Under the latter conditions, the simpler Double Quantification Rule (**Section 6.2.2**.) might be used with the GWPS set at a quantification limit. However, this could pose some interpretation problems. Subsequent testing against the background GWPS at the same compliance well concentration levels causing the initial detection monitoring exceedance, might very likely result in further excursions above the background GWPS. The more realistic option would be to collect and use additional compliance well data to establish a specific minimum intrawell background, and only apply the Double Quantification Rule at other wells not exhibiting detections. Even this approach might be unnecessarily stringent if a contaminant plume were to expand in size and gradually affect other compliance wells (now subject to GWPS testing).

CUMULATIVE & INDIVIDUAL TEST FALSE POSITIVE RATES

Each of the independent two-sample tests against background standards will have a roughly equal probability of being exceeded by chance alone. Since an exceedance in the compliance monitoring mode based on background can result in a need for corrective action, it is recommended that the individual test false positive rate be set sufficiently low. Much of the discussion in **Chapter 6**, **Section 6.2.2** is relevant here. An *a priori,* cumulative error design rate must first be identified. To allow for application of the Unified Guidance detection monitoring strategies and **Appendix D** tables, it is suggested that the .1 *SWFPR* value also be applied to two-sample background GWPS testing. In similar fashion to **Chapter 6** and **Part III**, this can be translated into individual test configurations.

If the single-sample confidence interval option will be used with an elevated GWPS, the compliance level test will have a very low probability of being exceeded by truly background data. Cumulative false positive error considerations are generally negligible. For testing compliance/assessment or corrective action hypotheses, there is still a need to identify an appropriately low single test false positive rate which meets the regulatory goals. Generally, a single test false positive error rate of .1 to .05 will be suitable with the recommended approach for defining the background GWPS.

UNIFIED GUIDANCE RECOMMENDATIONS

Two-Sample GWPS Definition and Testing

As indicated above, any of the detection monitoring tests described in **Chapter 6** might be selected for two- or multiple- sample background compliance testing. One highly recommended statistical test approach is a prediction limit. Either a parametric prediction limit for a future mean (**Section 18.2.2**) or a non-parametric prediction limit for a future median (**Section 18.3.2**) can be used, depending on the constituent being tested and its statistical and distributional characteristics (*e.g.*, detection rate, normality, *etc*.). It would be equally possible to utilize one of the 1-of-*m* future value prediction limit tests, on an interwell or intrawell basis. Use of repeat samples as part of the selected test is appropriate, although the expected number of annual compliance/corrective action samples may dictate which tests can apply.

One parametric example is the 1-of-1 future mean test. If the background data can be normalized, background observations are used to construct a parametric prediction limit with (1–α) confidence around a mean of order *p*, using the equation:

$$PL = \bar{x} + t_{1-\alpha,n-1} \cdot s \cdot \sqrt{\frac{1}{p} + \frac{1}{n}} \qquad [7.5]$$

The next *p* measurements from each compliance well are averaged and the future mean compared to the background prediction limit, *PL* (considered the background GWPS). In compliance/assessment monitoring, if any of the means exceeds the limit, those well-constituent pairs are deemed to be out of compliance. In corrective action, if the future mean is no greater than *PL*, it can be concluded that the well-constituent pair is sufficiently similar to background to be within the remediation goal. In both monitoring phases, the prediction limit is constructed to represent a reasonable upper limit on the

background distribution. Compliance point means above this limit are statistically different from background; means below it are similar to background.

If the background sample cannot be normalized perhaps due to a large fraction of non-detects, two-sample non-parametric upper prediction limit detection monitoring tests (**Chapters 18 & 19**) can be used. As an example, a maximal order statistic (often the highest or second-highest value) can be selected from background as a non-parametric 1-of-1 upper prediction limit test of the median. **Table 18-2** is used to guide the choice based on background sample size (n) and the achievable confidence level (α). The median of the next 3 measurements from each compliance well is compared to the upper prediction limit. As with the parametric case in compliance/assessment, if any of the medians exceeds the limit, those well-constituent pairs would be considered out of compliance. In corrective action, well-constituent pairs with medians no greater than the background prediction limit would be considered as having met the standard.

If background measurements for a particular constituent are all non-detect, the GWPS should be set equal to the highest RL. In similar fashion to detection monitoring, 1-of-2 or 1-of-3 future value prediction limit tests can be applied (**Section 6.2.2** Double Quantification rule).

Single-Sample GWPS Definition and Testing

For single-sample testing, the Unified Guidance recommendation is to define a fixed GWPS or ACL based on a background *upper tolerance limit* with 95% confidence and 95% coverage (**Chapter 17**). For normal background, the appropriate formula for the GWPS would be the same as that given in **Section 17.2.1**, namely:

$$GWPS = \bar{x} + \tau(n,.95,.95)\cdot s \qquad [7.6]$$

where n = number of background measurements, $\bar{x}$ and s represent the background sample mean and standard deviation, and τ is a tolerance factor selected from **Table 17-3**. If the background sample is a mixture of detects and non-detects, but the non-detect fraction is no more than 50%, a censored estimation method such as Kaplan-Meier or robust regression on order statistics [ROS] (**Chapter 15**) can be attempted to compute adjusted estimates of the background mean μ and standard deviation σ in equation [7.5].

For larger fractions of non-detects, a non-parametric tolerance limit can be constructed, as explained in **Section 17.2.2**. In this case, the GWPS median will often be set to the largest or second-largest observed value in background. **Table 17-4** can be used to determine the achieved confidence level ($1-\alpha$) associated with a 95% coverage GWPS constructed in this way. Ideally, enough background measurements should be used to set the tolerance limit as close to the target of 95% coverage, 95% confidence as possible. However, this could require very large background sample sizes ($n \geq 60$).

Multiple independent measurements are used to form either a mean or median confidence interval for comparison with the background GWPS. Preferably at least 4 distinct compliance point measurements should be used to define the mean confidence interval in the parametric case, and 3-7 values should be used with a non-parametric median test. The guidance does not recommend retesting in single-sample background GWPS compliance/assessment monitoring. An implicit kind of retesting is built in to any test of a sample mean or median as explained in **Section 19.3.2.**

In essence, the background tolerance limit is used to set a somewhat higher mean target GWPS which can accommodate both compliance and corrective action testing under background conditions. The GWPS in equation [7.6] can be interpreted as an approximation to the upper 95th percentile of the background distribution. It is designed to be a reasonable maximum on the likely range of background concentrations. It is high enough that compliance wells exceeding the GWPS via a confidence interval test (*i.e.*, LCL > GWPS) are probably impacted and not mere false positives. At the same time, successful remedial efforts must show that concentrations at contaminated wells have decreased to levels similar to background. The GWPS above represents an upper bound on background but is not so low as to make proof of remediation via an upper confidence limit [GWPS] impossible.

To ensure that the GWPS in equation [7.6] sets a reasonable target, the Unified Guidance recommends that at least 8 to 10 background measurements (*n*) be utilized, and more if available. If the background sample is not normal, but can be normalized via a transformation, the tolerance limit should be computed on the transformed measurements and the result ***back-transformed*** to obtain a limit in the concentration scale (see **Chapter 17** for further details).

TRADEOFFS IN BACKGROUND GWPS TESTING METHODS

A two-sample GWPS approach offers a stricter test of background exceedances. There is also greater flexibility in designing tests for a variety of future comparison values (single with repeat, small sample means, etc.). The true test parameter is explicitly defined by the type of test chosen. Non-parametric upper prediction limit tests also allow for greater flexibility when data sets include significant non-detect values or are not transformable to a normal distribution assumption. The approach suggested in this section accounts for the cumulative false positive error rate.

One negative feature of two-sample GWPS testing is that the test hypotheses cannot be reversed for correction action monitoring. The trigger for compliance/assessment testing may also be quite small, resulting in important consequences (the need to move to corrective action). It may also be difficult to demonstrate longer-term compliance following remedial activities, if the actual background is somewhat elevated.

Single-sample GWPS testing, by contrast, does allow for the reversal of test hypotheses. Using a suitable definition of the somewhat elevated GWPS takes into account background sample variability and size. Cumulative false positive error rates for compliance or corrective action testing are not considered, and standardized alpha error levels (.1 or .05) can be used. Exceedances under compliance monitoring also offer clear evidence of a considerable increase above background.

But applying an arbitrary increase above background recommended for single-sample testing may conflict with State anti-degradation policy. Defining the GWPS as a specific population parameter is also somewhat arbitrary. Using the suggested guidance approach for defining the GWPS in equation [7.6] above, may result in very high values if the data are not normal (including logarithmic or non-parametric applications). There is also less flexibility in identifying testing options, especially with data sets containing significant non-detect values. Annual testing with quarterly sampling may be the only realistic choice.

A possible compromise might utilize both approaches. That is, initially apply the two-sample approach for compliance/assessment testing. Then evaluate the single-sample approach with reversed

hypotheses. Some of the initially significant increases under the two-sample approach may also meet the upper confidence level limit when tested against the higher GWPS. Those well constituents that cannot meet this limit can then be subjected to corrective action remediation and full post-treatment testing. This implies that the background GWPS would be a range based on the two testing methods rather than a single value.

►EXAMPLE 7-1

A facility has triggered a significant increase under detection monitoring. One hazardous constituent (arsenic) was identified which must be tested against a background GWPS at six different compliance wells, since background well levels were above the appropriate arsenic MCL of 10 ug/l. Two semi-annual tests are required for compliance/assessment monitoring. Assume that arsenic had been detected in both background and downgradient wells, but was significantly higher in one of the compliance wells. It must be determined whether any of the compliance wells have exceeded their background GWPS, and might require corrective action.

Design a background GWPS monitoring system for the following arsenic data from the elevated Well #1, consisting of eight hypothetical historical intrawell background samples and four future annual values for two different simulated data distribution cases shown in the table below. Sample means and standard deviations are provided in the bottom row:

Compliance Well #1 Arsenic (μg/l)

Historical Well Data		Case 1	Case2
74.1	41.5	61.5	95.0
10.8	41.0	58.7	73.4
32.8	30.8	76.8	73.3
25.0	40.0	81.3	90.0
$\bar{x}$ = 37.0		$\bar{x}$ = 69.58	$\bar{x}$ = 82.93
s = 18.16		s = 11.15	s = 11.24

Background values were randomly generated from a normal distribution with a true mean of μ = 40 and a population standard deviation of σ = 16. Case 1 future data were from a normal distribution with a mean 1.5 times higher, while Case 2 data were from a normal distribution twice as high as the background true mean. Both cases used the same background population standard deviation. The intent of these simulated values is to allow exploration of both of the Unified Guidance recommended background GWPS methods when background increases are relatively modest and sample sizes small.

The two-sample background GWPS approach is first evaluated. Assume that the background data are normal and stationary (no evidence of spatial or temporal variation and other forms of statistical dependence). Given a likely limit of future quarterly sampling and required semi-annual evaluations, two guidance prediction limit options would seem appropriate—either a 1-of-2 future values or a 1-of-1 future mean size 2 test conducted twice a year. The 1-of-2 future values option is chosen.

Since there are a total of 6 compliance wells, one background constituent and two annual evaluations, there are a total of 12 annual background tests to be conducted. Either the Unified Guidance tables in **Appendix D** or R-script can be used to identify the appropriate prediction limit κ-

factor. For the 1-of-2 future values test, κ = **1.83** (found by interpolation from the second table on page D-118), based on w = 6, COC = 1, and two tests per year. The calculated prediction limit using the background data set statistics and κ-factor is 70.2 µg/l, serving as the background GWPS.

When the future values from the table above are tested against the GWPS, the following results are obtained. A "Pass" indicates that the compliance/assessment null hypothesis was achieved, while a "Fail" indicates that the alternative hypothesis (the GWPS has been exceeded) is accepted.

Well #1 As Compliance Comparisons
1-of-2 Future Values Test (µg/l)

Case 1 (data)	Result	Case 2 (data)	Result
61.5		95.0	
58.7	Pass	73.4	Fail
76.8		73.3	
81.3	Fail	90.0	Fail

GWPS = 70.2

Both cases indicate at least one GWPS exceedance using the 1-of-2 future values tests. These may be indications of a statistically significant increase above background, but the outcome for Case 1 is somewhat troubling. While a 50% increase above background (based on the simulated population parameters) is potentially significant, more detailed power evaluations indicate that such a detected exceedance would only be expected about 24% of the time (using R-script power calculations with a Z-value of 1.25 standard deviations above background for the 1-of-2 future values test). In contrast, the 2.5 Z-value for Case 2 would be expected to be exceeded about 76% of the time. In order to further evaluate the extent of significance of these results, the single-sample GWPS method is also considered.

Following the guidance above, define the single-sample mean GWPS using equation [7.6] for the upper 95% confidence, 95% proportion tolerance limit. Then apply upper and lower normal mean confidence intervals tests of the Case 1 and 2 n = 4 sample data using equations [7.3] and [7.4].

From **Table 21-9** on page D-246, a τ-factor of 3.187 is used with the background mean and standard deviation to generate the GWPS = 94.9. One-way upper and lower mean confidence levels are evaluated at 90 or 95% confidence for the tests and compared to the fixed background GWPS.

LCL test Pass/Fail results are the same as above for the two-sample compliance test. However, a "Pass" for the UCL test implies that the alternative hypothesis (less than the standard) is accepted while a "Fail" implies greater than or equal to the GWPS under corrective action monitoring hypotheses:

As Mean Confidence Interval Tests Against Background GWPS (μg/l)

LCL Test				UCL Test			
90% LCL	Result	95% LCL	Result	90% UCL	Result	95% UCL	Result
			Case 1 Data				
60.5	Pass	56.5	Pass	78.7	Pass	82.7	Pass
			Case 2 Data				
73.7	Pass	69.7	Pass	92.1	Pass	96.2	Fail

GWPS = 94.9

For either chosen significance level, the Case 1 90% and 95% UCLs of 78.7 and 82.7 are below the GWPS and the alternative corrective action hypothesis (the mean is less than the standard) can be accepted. For Case 2, the 90% UCL of 92.1 is below the GWPS, but the 95% UCL of 96.2 is above. If a higher level of test confidence is appropriate, the Case 2 arsenic values can be considered indicative of the need for corrective action.

If only the single-sample background GWPS approach were applied to the same data as above in compliance/assessment monitoring tests, neither case mean LCLs would exceed the standard, and no corrective action monitoring would be necessary. However, it should be noted from the example that this approach does allow for a significant increase above the reference background level before any action would be indicated. ◄

The approaches provided above presume that well constituent data subject to background GWPS testing are *stationary* over time. If sampling data show evidence of a trend, the situation becomes more complicated in making compliance or corrective action test decisions. Two- and single-sample stationary scenarios for identifying standards may not be appropriate. Trend behavior can be determined by applying one of the methods provided in **Chapter 17** (e.g., linear regression or Mann-Kendall trend tests) to historical data. A significant increasing slope can be indicative of a background exceedance, although it should be clear that the increase is not due to natural conditions. A decreasing or non-significant slope can be considered evidence for compliance with historical background. The most problematic standard would be setting an eventual background target for compliance testing under corrective action. To a great extent, it will depend on site-specific conditions including the behavior of specific constituent subject to remediation. A background GWPS might be determined following the period of remediation and monitoring when aquifer conditions have hopefully stabilized.

Setting and applying background GWPSs have not received a great deal of attention in previous guidance. The discussions and example above help illustrate the somewhat difficult regulatory choices that need to be made. A regulatory agency needs to determine what levels, if any, above background can be considered acceptable. A further consideration is the degree of importance placed on background GWPS exceedances, particularly when tested along with constituents having health-based limits. Existing regulatory programs may have already developed procedures to deal with many of the issues discussed in this section.

CHAPTER 8. SUMMARY OF RECOMMENDED METHODS

This chapter provides a quick guide to the statistical procedures discussed within the Unified Guidance. The first section is a basic road map designed to encourage the user to ask a series of key questions. The other sections offer thumbnail sketches of each method and a matrix of options to help in selecting the right procedure, depending on site-specific characteristics and constraints.

8.1 SELECTING THE RIGHT STATISTICAL METHODS

Choosing appropriate statistical methods is important in developing a sound groundwater monitoring statistical program. The statistical test(s) should be selected to match basic site-specific characteristics such as number and configuration of wells, the water quality constituents being measured, and general hydrology. Statistical methods should also be selected with reference to the statistical characteristics of the monitored parameters — proportion of non-detects, type of concentration distribution (***e.g.***, normal, lognormal), presence or absence of spatial variability, ***etc.***

Because site conditions and permit requirements vary considerably, no single "cookbook" approach is readily available to select the right statistical method. The best strategy is to consider site-specific conditions and ask a series of questions. A table of recommended options (**Table 8-1**) and summary descriptions is presented in **Section 8.2** to help select an appropriate basic approach.

The first question is: what stage of monitoring is required? Detection monitoring is the first stage of any groundwater monitoring program and typically involves comparisons between measurements of background and compliance point groundwater. Most of the methods described in this document (***e.g.***, prediction limits, control charts, tests for trend, ***etc.***) are designed for facilities engaged in detection monitoring. However, it must be determined whether an interwell (***e.g.***, upgradient-to-downgradient) or an intrawell test is warranted. This entails consideration of the site hydrology, constituent detection rates, and deciding whether separate (upgradient) wells or past intrawell data serves as the most appropriate and representative background.

Compliance/assessment monitoring is required for facilities that no longer meet the requirements of a detection monitoring program by exhibiting statistically significant indications of a release to groundwater. Once in compliance/assessment, compliance point measurements are typically tested against a fixed GWPS. Examples of fixed standards include Maximum Concentration Limits [MCL], risk-derived limits or a single limit derived from background data. The most appropriate statistical method for tests against GWPS is a lower confidence limit. The type of confidence limit will depend on whether the regulatory standard represents an average concentration; an absolute maximum, ceiling, or upper percentile; or whether the compliance data exhibit a trend over time.

In cases where no fixed GWPS is specified for a particular constituent, compliance point data may be directly compared against background data. In this situation, the most appropriate statistical method is

one or another detection monitoring two- or multiple-sample tests using the critical design limit as the GWPS (discussed in **Section 7.5**).

Corrective action is reserved for facilities where evidence of a groundwater release is confirmed above a GWPS. In these situations, the facility is required to submit an appropriate remediation plan to the Regional Administrator and to institute steps to insure adequate containment and/or clean-up of the release. Remediation of groundwater can be very costly and also difficult to measure. EPA has not adopted a uniform approach in the setting of clean-up standards or how one should determine whether those clean-up standards have been attained. Some guidance on this issue is given in the EPA document, ***Methods for Evaluating the Attainment of Cleanup Standards, Volume II: Groundwater*** (EPA, 1992).

The null hypothesis in corrective action testing is reversed from that of detection and compliance/assessment monitoring. Not only is it assumed that contamination is above the compliance or clean-up standard, but corrective action should continue until the average concentration level is below the clean-up limit for periods specified in the regulations. For any fixed-value standard (e.g., the GWPS or a remediation goal) a reasonable and consistent statistical test for corrective action is an *upper* confidence limit. The type of confidence limit will depend on whether the data have a stable mean concentration or exhibit a trend over time. For those well constituents requiring remediation, there will be a period of activity before formal testing can take place. A number of statistical techniques (e.g. trend testing) can be applied to the data collected in this interim period to gauge prospects for eventual GWPS compliance. **Section 7.5** describes corrective action testing limitations involving a two-sample GWPS.

Another major question involves the statistical distribution most appropriate to the observed measurements. Parametric tests are those which assume the underlying population follows a known and identifiable distribution, the most common examples in groundwater monitoring being the normal and the lognormal. If a specific distribution cannot be determined, non-parametric test methods can be used. Non-parametric tests do not require a known statistical distribution and can be helpful when the data contain a substantial proportion of non-detects. All of the parametric tests described in the Unified Guidance, except for control charts, have non-parametric counterparts that can be used when the underlying distribution is uncertain or difficult to test.

A special consideration in fitting distributions is the presence of non-detects, also known as left-censored measurements. As long as a sample contains a small fraction of non-detects (*i.e.*, no more than 10-15%), simple substitution of half the reporting limit [RL] is generally adequate. If the proportion of non-detects is substantial, it may be difficult or impossible to determine whether a specific parametric distributional model provides a good fit to the data. For some tests, such as the *t*-test, one can switch to a non-parametric test with little loss of power or accuracy. Non-parametric interval tests, however, such as prediction and tolerance limits, require substantially more data before providing statistical power equivalent to *parametric* intervals. Partly because of this drawback, the Unified Guidance discusses methods to adjust datasets with significant fractions of non-detects so that parametric distributional models may still be used (**Chapter 15**).

The Unified Guidance now recommends a single, consistent D*ouble Quantification rule* approach for handling constituents that have either never been detected or have not been recently detected. Such constituents are *not* included in cumulative annual *site-wide false positive error rate* [SWFPR] computations; and no special adjustment for non-detects is necessary. Any confirmed quantification (*i.e.*,

two consecutive detections above the RL) at a compliance point provides sufficient evidence of groundwater contamination by that parameter.

A key question when picking a test for detection monitoring is whether traditional background-to-downgradient interwell or single-well intrawell tests are appropriate. If intrawell testing is appropriate, historical measurements form the individual compliance well's own background while future values are tested against these data. Intrawell tests eliminate any natural spatial differences among monitoring wells. They can also be used when the groundwater flow gradient is uncertain or unstable, since all samples being tested are collected from the same well.

Possible disadvantages to intrawell tests also need to be considered. First, if the compliance well has already been impacted, intrawell background will also be impacted. Such contaminated background may provide a skewed comparison to later data from the same well, making it difficult to identify contaminated groundwater in the future. Secondly, if intrawell background is constructed from only a few early measurements, considerable time may be needed to accumulate a sufficient number of background observations (via periodic updating) to run a statistically powerful test.

If a compliance well has already been impacted by previous contamination, trend testing can still indicate whether conditions have deteriorated since intrawell background was collected. For sites historically contaminated above background, the only way to effectively monitor compliance wells may be to establish an historical intrawell baseline and measure increases above this baseline.

Besides trend tests, techniques recommended for intrawell comparisons include intrawell prediction limits, control charts, and sometimes the Wilcoxon rank-sum test. The best choice between these methods is not always clear. Since there is no non-parametric counterpart to control charts, the choice will depend on whether the data is normal or can be normalized via a transformation. New guidance for control charts shows they also can be designed to incorporate retesting. For sites with a large number of well-constituent pairs, intrawell prediction limits can incorporate retesting to meet specific site-wide false positive rate and statistical power characteristics. Parametric intrawell prediction limits can be used with background that is normal or transformable to normality; non-parametric versions can also be applied for many other data sets.

If interwell, upgradient-to-downgradient tests are appropriate, the choice of statistical method depends primarily on the number of compliance wells and constituents being monitored, the number of observations available from each of these wells, and the detection rates and distributional properties of these parameters. If a very small number of comparisons must be tested (*i.e.*, two or three compliance wells versus background, for one or two constituents), a *t*-test or Wilcoxon rank-sum test may be appropriate if there are a sufficient number of compliance measurements (*i.e.*, at least two per well).

For other cases, the Unified Guidance recommends a prediction limit or control chart constructed from background. Whenever more than a few statistical tests must be run, retesting should be incorporated into the procedure. If multiple observations per compliance well can be collected during a given evaluation period, either a prediction limit for 'future' observations, a prediction limit for means or medians, or a control chart can be considered, depending on which option best achieves statistical power and SWFPR targets, while balancing the site-specific costs and feasibility of sampling. If only one observation per compliance well can be collected per evaluation, the only practical choices are a prediction limit for individual observations or a control chart.

8.2 TABLE 8-1 INVENTORY OF RECOMMENDED METHODS

Chapter 9. Exploratory Tools		
Statistical Method	Chapter	Use
Time Series Plot	**§9.1**	Plot of measurement levels over time; Useful for assessing trends, data inconsistencies, etc.
Box Plot	**§9.2**	Graphical summary of sample distribution; Useful for comparing key statistical characteristics in multiple wells
Histogram	**§9.3**	Graphical summary of sample distribution; Useful for assessing probability density of single data set
Scatter Plot	**§9.4**	Diagnostic tool; Plot of one variable vs. another; Useful for exploring statistical associations
Probability Plot	**§9.5**	Graphical fit to normality; Useful for raw or transformed data
Chapter 10. Fitting Distributions		
Statistical Method	Chapter	Use
Skewness Coefficient	**§10.4**	Measures symmetry/asymmetry in distribution; Screening level test for plausibility of normal fit
Coefficient of Variation	**§10.4**	Measures symmetry/asymmetry in distribution; Screening tool for plausibility of normal fit; Only for non-negative data
Shapiro-Wilk Test	**§10.5.1**	Numerical normality test of a single sample; for $n \leq 50$
Shapiro-Francía Test	**§10.5.2**	Numerical test of normality for a single sample; Supplement to Shapiro-Wilk; Use with $n > 50$
Filliben's Probability Plot Correlation Coefficient	**§10.6**	Numerical test of normality for a single sample; Interchangeable with Shapiro-Wilk; Use with $n \leq 100$; Good supplement to probability plot
Shapiro-Wilk Multiple Group Test	**§10.7**	Extension of Shapiro-Wilk test for multiple samples with possibly different means and/or variances; Good check to use with Welch's *t*-test
Chapter 11. Equality of Variance		
Statistical Method	Chapter	Use
Box Plots (side-by-side)	**§11.1**	Graphical test of differences in population variances; Good screening tool for equal variance assumption in ANOVA
Levene's Test	**§11.2**	Numerical, robust ANOVA-type test of equality of variance for ≥ 2 populations; Useful for testing assumptions in ANOVA
Mean-SD Scatter Plot	**§11.3**	Visual test of association between SD and mean levels across group of wells; Use to check for proportional effect or if variance-stabilizing transformation is needed
Chapter 12. Outliers		
Statistical Method	Chapter	Use
Probability Plot	**§12.1**	Graphical fit of distribution to normality; Useful for identifying extreme points not coinciding with predicted tail of distribution
Box Plot	**§12.2**	Graphical screening tool for outliers; quasi-non-parametric, only requires rough symmetry in distribution
Dixon's Test	**§12.3**	Numerical test for single low or single high outlier; Use when $n \leq 25$
Rosner's Test	**§12.4**	Numerical test for up to 5 outliers in single dataset; Recommended when $n \geq 20$; User must identify a specific number of possible outliers before running

Chapter 13. Spatial Variation		
Statistical Method	Chapter	Use
Box Plots (side-by-side)	**§13.2.1**	Quick screen for spatial variability; Look for noticeably staggered boxes
One-Way Analysis of Variance [ANOVA] for Spatial Variation	**§13.2.2**	Test to compare means of several populations; Use to identify spatial variability across a group of wells and to estimate pooled (background) standard deviation for use in intrawell tests; Data must be normal or normalized; Assumption of equal variances across populations
Chapter 14. Temporal Variability		
Statistical Method	Chapter	Use
Time Series Plot (parallel)	**§14.2.1**	Quick screen for temporal (and/or spatial) variation; Look for parallel movement in the graph traces at several wells over time
One-way ANOVA for Temporal Effects	**§14.2.2**	Test to compare means of distinct sampling events, in order to assess systematic temporal dependence across wells; Use to get better estimate of (background) variance and degrees of freedom in data with temporal patterns; Residuals from ANOVA also used to create stationary, adjusted data
Sample Autocorrelation Function	**§14.2.3**	Plot of autocorrelation by lag between sampling events; Requires approximately normal data; Use to test for temporal correlation and/or to adjust sampling frequency
Rank von Neumann Ratio	**§14.2.4**	Non-parametric numerical test of dependence in time-ordered data series; Use to test for first-order autocorrelation in data from single well or population
Darcy Equation	**§14.3.2**	Method to approximate groundwater flow velocity; Use to determine sampling interval guaranteeing physical independence of consecutive groundwater samples; Does not ensure statistical independence
Seasonal Adjustment (single well)	**§14.3.3**	Method to adjust single data series exhibiting seasonal correlations (*i.e.*, cyclical fluctuations); At least 3 seasonal cycles must be evident on time series plot
Temporally-Adjusted Data Using ANOVA	**§14.3.3**	Method to adjust multiple wells for a common temporal dependence; Use adjusted data in subsequent tests
Seasonal Mann-Kendall Test	**§14.3.4**	Extension of Mann-Kendall trend test when seasonality is present; At least 3 seasonal cycles must be evident
Chapter 15. Managing Non-Detect Data		
Statistical Method	Chapter	Use
Simple Substitution	**§15.2**	Simplest imputation scheme for non-detects; Useful when ≤ 10-15% of dataset is non-detect
Censored Probability Plot	**§15.3**	Probability plot for mixture of non-detects and detects; Use to check normality of left-censored sample
Kaplan-Meier	**§15.3**	Method to estimate mean and standard deviation of left-censored sample; Use when ≤ 50% of dataset is non-detect; Multiple detects and non-detects must originate from same distribution
Robust Regression on Order Statistics	**§15.4**	Method to estimate mean and standard deviation of left-censored sample; Use when ≤ 50% of dataset is non-detect; Multiple detects and non-detects must originate from same distribution
Cohen' Method and Parametric Regression on Order Statistics	**§15.5**	Other methods to estimate mean and standard deviation of left-censored sample; Use when ≤ 50% of dataset is non-detect; Detects and non-detects must originate from same distribution and there must be a single censoring limit

Chapter 16. Two-sample Tests		
Statistical Method	Chapter	Use
Pooled Variance *t*-Test	**§16.1.1**	Test to compare means of two populations; Data must be normal or normalized, with no significant spatial variability; Useful at very small sites in upgradient-to-downgradient comparisons; Also useful for updating background; Population variances must be equal
Welch's *t*-Test	**§16.1.2**	Test to compare means of two populations; Data must be normal or normalized, with no significant spatial variability; Useful at very small sites in interwell comparisons; Also useful for updating background; Population variances can differ
Wilcoxon Rank-Sum Test	**§16.2**	Non-parametric test to compare medians of two populations; Data need not be normal; Some non-detects OK; Should have no significant spatial variability; Useful at very small sites in interwell comparisons and for certain intrawell comparisons; Also useful for updating background
Tarone-Ware Test	**§16.3**	Extension of Wilcoxon rank-sum; non-parametric test to compare medians of two populations; Data need not be normal; Designed to accommodate left-censored data; Should have no significant spatial variability; Useful at very small sites in interwell comparisons and for certain intrawell comparisons; Also useful for updating background
Chapter 17. ANOVA, Tolerance Limits, & Trend Tests		
Statistical Method	Chapter	Use
One-Way ANOVA	**§17.1.1**	Test to compare means across multiple populations; Data must be normal or normalized; Should have no significant spatial variability if used as interwell test; Assumes equal variances; Mandated in some permits, but generally superceded by other tests; Useful for identifying spatial variation; RMSE from ANOVA can be used to improve intrawell background limits
Kruskal-Wallis Test	**§17.1.2**	Test to compare medians across multiple populations; Data need not be normal; some non-detects OK; Should have no significant spatial variability if used as interwell test; Useful alternative to ANOVA for identifying spatial variation
Tolerance Limit	**§17.2.1**	Test to compare background vs. ≥ 1 compliance well; Data must be normal or normalized; Should have no significant spatial variability if used as interwell test; Alternative to ANOVA; Mostly superceded by prediction limits; Useful for constructing alternate clean-up standard in corrective action
Non-parametric Tolerance Limit	**§17.2.2**	Test to compare background vs. ≥ 1 compliance well; Data need not be normal; Non-Detects OK; Should have no significant spatial variability if used as interwell test; Alternative to Kruskal-Wallis; Mostly superceded by prediction limits
Linear Regression	**§17.3.1**	Parametric estimate of linear trend; Trend residuals must be normal or normalized; Useful for testing trends in background or at already contaminated wells; Can be used to estimate linear association between two random variables
Mann-Kendall Trend Test	**§17.3.2**	Non-parametric test for linear trend; Non-detects OK; Useful for documenting upward trend at already contaminated wells or where trend already exists in background
Theil-Sen Trend Line	**§17.3.3**	Non-parametric estimate of linear trend; Non-detects OK; Useful for estimating magnitude of an increasing trend in conjunction with Mann-Kendall test

Chapter 18. Prediction Limit Primer		
Statistical Method	Chapter	Use
Prediction Limit for m Future Values	**§18.2.1**	Test to compare m measurements from compliance well against background; Data must be normal normalized; Useful in retesting schemes; Can be adapted to either intrawell or interwell tests; No significant spatial variability allowed if used as interwell test
Prediction Limit for Future Mean	**§18.2.2**	Test to compare mean of compliance well against background; Data must be normal or normalized; Useful alternative to traditional ANOVA; Can be useful in retesting schemes; Most useful for interwell (*e.g.*, upgradient to downgradient) comparisons; No significant spatial variability allowed if used as interwell test
Non-Parametric Prediction Limit for m Future Values	**§18.3.1**	Non-parametric test to compare m measurements from compliance well against order statistics of background; Non-normal data and/or non-detects OK; Useful in non-parametric retesting schemes; Should have no significant spatial variability if used as interwell test
Non-parametric Prediction Limit for Future Median	**§18.3.2**	Test to compare median of compliance well against order statistics of background; Non-normal data and/or non-detects OK; Useful in non-parametric retesting schemes; Most useful for interwell (*e.g.*, upgradient to downgradient) comparisons; No significant spatial variability allowed if used as interwell test
Chapter 19. Prediction Limit Strategies with Retesting		
Statistical Method	Chapter	Use
Prediction Limits for Individual Observations With Retesting	**§19.3.1**	Tests individual compliance point measurements against background; Data must be normal or normalized; Assumes common population variance across wells; No significant spatial variability allowed if used as interwell test; Replacement for traditional ANOVA, extends Dunnett's multiple comparison with control (MCC) procedure; Allows control of SWFPR across multiple well-constituent pairs; Retesting explicitly incorporated; Useful at any size site
Prediction Limits for Means With Retesting	**§19.3.2**	Tests compliance point means against background; Data must be normal or normalized; Assumes common population variance across wells; No significant spatial variability allowed if used as interwell test; Replacement for traditional ANOVA, extends Dunnett's multiple comparison with control (MCC) procedure; More flexible than a series of intrawell t-tests if used as intrawell test; Allows control of SWFPR across multiple well-constituent pairs; Must be feasible to collect ≥2 resamples per evaluation period to incorporate retesting; 1-of-1 scheme does not require explicit retesting
Non-Parametric Prediction Limits for Individual Observations With Retesting	**§19.4.1**	Non-parametric test of individual compliance point observations against background; Non-normal data and/or non-detects OK; No significant spatial variability allowed if used as interwell test; Retesting explicitly incorporated; Large background sample size helpful
Non-Parametric Prediction Limits for Medians With Retesting	**§19.4.2**	Non-parametric test of compliance point medians against background; Non-normal and/or non-detects OK; No significant spatial variability allowed if used as interwell test; Large background sample size helpful; Must be feasible to collect ≥ 3 resamples per evaluation period to incorporate retesting; 1-of-1 scheme does not require explicit retesting

Chapter 20. Control Charts		
Statistical Method	Chapter	Use
Shewhart-CUSUM Control Chart	**§20.2**	Graphical test of significant increase above background; Data must be normal or normalized; Some non-detects OK if left-censored adjustment made; At least 8 background observations recommended; Viable alternative to prediction limits; Retesting can be explicitly incorporated; Control limits can be set via published literature or Monte Carlo simulation
Chapter 21. Confidence Intervals		
Statistical Method	Chapter	Use
Confidence Interval Around Normal Mean	**§21.1.1**	Data must be normal; Some non-detects OK if left-censored adjustment made; Used in compliance/assessment or corrective action to compare compliance well against fixed, mean-based groundwater standard; Should be no significant trend; 4 or more observations recommended
Confidence Interval Around Lognormal Geometric Mean	**§21.1.2**	Data must be lognormal; Some non-detects OK if left-censored adjustment made; Used in compliance/assessment or corrective action to compare compliance well against fixed, mean-based groundwater standard; Should be no significant trend; 4 or more observations recommended; Geometric mean equivalent to lognormal median, smaller than lognormal mean
Confidence Interval Around Lognormal Arithmetic Mean	**§21.1.3**	Data must be lognormal; Some non-detects OK if left-censored adjustment made; Used in compliance/assessment or corrective action to compare compliance well against fixed, mean-based groundwater standard; Should be no significant trend; 4 or more observations recommended; Lognormal arithmetic mean larger than lognormal geometric mean
Confidence Interval Around Upper Percentile	**§21.1.4**	Data must be normal or normalized; Some non-detects OK if left-censored adjustment made; Used in compliance/assessment to compare compliance well against percentile-based or maximum groundwater standard; Should be no significant trend
Non-Parametric Confidence Interval around Median	**§21.2**	For non-normal, non-lognormal data; Non-detects OK; Used in compliance/assessment or corrective action to compare compliance well against fixed, mean-based groundwater standard; Should be no significant trend; 7 or more observations recommended
Non-Parametric Confidence Interval Around Upper Percentile	**§21.2**	For non-normal, non-lognormal data; Non-detects OK; Used in compliance/assessment or corrective action to compare compliance well against percentile-based or maximum groundwater standard; Should be no significant trend; Large background sample size helpful
Confidence Band Around Linear Regression	**§21.3.1**	Use on data with significant trend; Trend residuals must be normal or normalized; Used in compliance/assessment or corrective action to compare compliance well against fixed groundwater standard; ≥ 8 observations recommended
Non-parametric Confidence Band Around Theil-Sen Line	**§21.3.2**	Use on data with significant trend; Non-normal data and/or non-detects OK; Used in compliance/assessment or corrective action to compare compliance well against fixed groundwater standard; Bootstrapping of Theil-Sen trend line used to construct confidence band

8.3 METHOD SUMMARIES

TIME SERIES PLOT (SECTIONS 9.1 AND 14.2.1)

Basic purpose: Diagnostic and exploratory tool. It is a graphical technique to display changes in concentrations at one or more wells over a specified period of time or series of sampling events.

Hypothesis tested: Not a formal statistical test. Time series plots can be used to informally gauge the presence of temporal and/or spatial variability in a collection of distinct wells sampled during the same time frame.

Underlying assumptions: None.

When to use: Given a collection of wells with several sampling events recorded at each well, a time series plot can provide information not only on whether the mean concentration level changes from well to well (an indication of possible spatial variation), but also on whether there exists time-related or temporal dependence in the data. Such temporal dependence can be seen in parallel movement on the time series plot, that is, when several wells exhibit the same pattern of up-and-down fluctuations over time.

Steps involved: 1) For each well, make a plot of concentration against time or date of sampling for the sampling events that occurred during the specified time period; 2) Make sure each well is identified on the plot with a distinct symbol and/or connected line pattern (or trace); 3) To observe possible spatial variation, look for well traces that are substantially separated from one another in concentration level; 4) To look for temporal dependence, look for well traces that rise and fall together in roughly the same (parallel) pattern; 5) To ensure that artificial trends due to changing reporting limits are not reported, plot any non-detects with a distinct symbol, color, and/or fill.

Advantages/Disadvantages: Time series plots are an excellent tool for examining the behavior of one or more samples over time. Although, they do not offer the compact summary of distributional characteristics that, say, box plots do, time series plots display each and every data point and provide an excellent initial indication of temporal dependence. Since temporal dependence affects the underlying variability in the data, its identification is important so adjustments can be made to the estimated standard deviation.

BOX PLOT (SECTIONS 9.2, 12.2, AND 13.2.1)

Basic purpose: Diagnostic and exploratory tool. Graphical summary of data distribution; gives compact picture of central tendency and dispersion.

Hypothesis tested: Although not a formal statistical test, a side-by-side box plot of multiple datasets can be used as a rough indicator of either unequal variances or spatial variation (via unequal means/medians). Also serves as a quasi-non-parametric screening tool for outliers in a symmetric population.

Underlying assumptions: When used to screen outliers, underlying population should be approximately symmetric.

When to use: Can be used as a quick screen in testing for unequal variances across multiple populations. Box lengths indicate the range of the central 50% of sample data values. Substantially different box lengths suggest possibly different population variances. It is useful as a rough indication of spatial variability across multiple well locations. Since the median (and often the mean) are graphed on each box, significantly staggered medians and/or means on a multiple side-by-side box plot can suggest possibly different population means at distinct well locations. Can also be used to screen for outliers: values falling beyond the 'whiskers' on the box plot are labeled as potential outliers.

Steps involved: 1) Compute the median, mean, lower and upper quartiles (*i.e.*, 25th and 75th percentiles) of each dataset; 2) Graph each set of summary statistics side-by-side on the same set of axes. Connect the lower and upper quartiles as the ends of a box, cut the box in two with a line at the median, and use an 'X' or other symbol to represent the mean. 3) Compute the 'whiskers' by extending lines below and above the box by an amount equal to 1.5 times the interquartile range [IQR].

Advantages/Disadvantages: The box plot is an excellent screening tool and visual aid in diagnosing either unequal variances for testing the assumptions of ANOVA, the possible presence of spatial variability, or potential outliers. It is not a formal statistical test, however, and should generally be used in conjunction with numerical test procedures.

HISTOGRAM (SECTION 9.3)

Basic purpose: Diagnostic and exploratory tool. It is a graphical summary of an entire data distribution.

Hypothesis tested: Not a formal statistical test.

Underlying assumptions: None.

When to use: Can be used as a rough estimate of the probability density of a single sample. Shape of histogram helps determine whether the distribution is symmetric or skewed. For larger data sets, histogram can be visually compared to a normal distribution or other known model to assess whether the shapes are similar.

Steps involved: 1) Sort and bin the data set into non-overlapping concentration segments that span the range of measurement values; 2) Create a bar chart of the bins created in **Step 1**: put the height of each bar equal to the number or fraction of values falling into each bin.

Advantages/Disadvantages: The histogram is a good visual aid in exploring possible distributional models that might be appropriate. Since it is not a formal test, there is no way to judge possible models solely on the basis of the histogram; however, it provides a visual 'feel' for a data set.

SCATTER PLOT (SECTION 9.4)

Basic purpose: Diagnostic tool. It is a graphical method to explore the association between two random variables or two paired statistical samples.

Hypothesis tested: None.

Underlying Assumptions: None.

When to use: Useful as an exploratory tool for discovering or identifying statistical relationships between pairs of variables. Graphically illustrates the degree of correlation or association between two quantities.

Steps involved: Using Cartesian pairs of the variables of interest, graph each pair on the scatter plot, using one symbol per pair.

Advantages/Disadvantages: A scatter plot is not a formal test, but rather an excellent exploratory tool. Helps identify statistical relationships.

PROBABILITY PLOT (SECTIONS 9.5 AND 12.1)

Basic purpose: Diagnostic tool. A graphical method to compare a dataset against a particular statistical distribution, usually the normal. Designed to show how well the data match up to or 'fit' the hypothesized distribution. An absolutely straight line fit indicates perfect consistency with the hypothesized model.

Hypothesis tested: Although not a formal test, the probability plot can be used to graphically indicate whether a dataset is normal. The straighter the plot, the more consistent the dataset with a null hypothesis of normality; significant curves, bends, or other non-linear patterns suggest a rejection of the normal model as a poor fit.

Underlying Assumptions: All observations come from a single statistical population.

When to use: Can be used as a graphical indication of normality on a set of raw measurements or, by first making a transformation, as an indication of normality on the transformed scale. It should generally be supplemented by a formal numerical test of normality. It can be used on the residuals from a one-way ANOVA to test the joint normality of the groups being compared. The test can also be used to help identify potential outliers (*i.e.*, individual values not part of the same basic underlying population).

Steps involved: 1) Order the dataset and determine matching percentiles (or quantiles) from the hypothesized distribution (typically the standard normal); 2) Plot the ordered data values against the matching percentiles; 3) Examine the plot for a straight line fit.

Advantages/Disadvantages: Not a formal test of normality; however, the probability plot is an excellent graphical supplement to any goodness-of-fit test. Because each data value is depicted, specific departures from normality can be identified (*e.g.*, excessive skewness, possible outliers, *etc.*).

SKEWNESS COEFFICIENT (SECTION 10.4)

Basic purpose: Diagnostic tool. Sample statistic designed to measure the degree of symmetry in a sample. Because the normal distribution is perfectly symmetric, the skewness coefficient can provide a quick indication of whether a given dataset is symmetric enough to be consistent with the normal model. Skewness coefficients close to zero are consistent with normality; skewness values large in absolute value suggest the underlying population is asymmetric and non-normal.

Hypothesis tested: The skewness coefficient is used in groundwater monitoring as a screening tool rather than a formal hypothesis test. Still, it can be used to roughly test whether a given sample is normal by using the following rule of thumb: if the skewness coefficient is no greater than one in absolute value, accept a null hypothesis of normality; if not, reject the normal model as ill-fitting.

Underlying Assumptions: None

Steps involved: 1) Compute skewness coefficient; 2) Compare to cutoff of 1; 3) If skewness is greater than 1, considering running a formal test of normality.

Advantages/Disadvantages: Fairly simple calculation, good screening tool. Skewness coefficient can be positive or negative, indicating positive or negative skewness in the dataset, respectively. Measures symmetry rather than normality, per se; since other non-normal distributions can also be symmetric, might give a misleading result. Not as powerful or accurate a test of normality as either the Shapiro-Wilk or Filliben tests, but a more accurate indicator than the coefficient of variation, particularly for data on a transformed scale.

COEFFICIENT OF VARIATION [CV] (SECTION 10.4)

Basic purpose: Diagnostic tool. Sample statistic used to measure skewness in a sample of positively-valued measurements. Because the CV of positively-valued normal measurements must be close to zero, the CV provides an easy indication of whether a given sample is symmetric enough to be normal. Coefficients of variation close to zero are consistent with normality; large CVs indicate a skewed, non-normal population.

Hypothesis tested: The coefficient of variation is not a formal hypothesis test. Still, it can be used to provide a 'quick and easy' gauge of non-normality: if the CV exceeds 0.5, the population is probably not normal.

Underlying Assumptions: Sample must be positively-valued for CV to have meaningful interpretation.

Steps involved: 1) Compute sample mean and standard deviation; 2) Divide standard deviation by mean to get coefficient of variation.

Advantages/Disadvantages: Simple calculation, good screening tool. It measures skewness and variability in positively-valued data. Not an accurate a test of normality, especially if data have been transformed.

SHAPIRO-WILK AND SHAPIRO-FRANCÍA TESTS (SECTION 10.5)

Basic purpose: Diagnostic tool and a formal numerical goodness-of-fit test of normality. Shapiro-Francía test is a close variant of the Shapiro-Wilk useful when the sample size is larger than 50.

Hypothesis tested: H_0 — the dataset being tested comes from an underlying normal population. H_A — the underlying population is non-normal (note that the form of this alternative population is not specified).

Underlying assumptions: All observations come from a single normal population.

When to use: To test normality on a set of raw measurements or following transformation of the data. It can also be used with the residuals from a one-way ANOVA to test the joint normality of the groups being compared.

Steps involved (for Shapiro-Wilk): 1) Order the dataset and compute successive differences between pairs of extreme values (*i.e.*, most extreme pair = maximum – minimum, next most extreme pair = 2nd largest – 2nd smallest, *etc.*); 2) Multiply the pair differences by the Shapiro-Wilk coefficients and compute the Shapiro-Wilk test statistic; 3) Compare the test statistic against an α-level critical point; 4) Values higher than the critical point are consistent with the null hypothesis of normality, while values lower than the critical point suggest a non-normal fit.

Advantages/Disadvantages: The Shapiro-Wilk procedure is considered one of the very best tests of normality. It is much more powerful than the skewness coefficient or chi-square goodness-of-fit test. The Shapiro-Wilk and Shapiro-Francía test statistics will tend to be large (and more indicative of normality) when a probability plot of the same data exhibits a close-to-linear pattern. Special Shapiro-Wilk coefficients are available for sample sizes up to 50. For larger sample sizes, the Shapiro-Francía test does not require a table of special coefficients, just the ability to compute inverse normal probabilities.

FILLIBEN'S PROBABILITY PLOT CORRELATION COEFFICIENT TEST (SECTION 10.6)

Basic purpose: Diagnostic tool and a formal numerical goodness-of-fit procedure to test for normality.

Hypothesis tested: H_0 — the dataset being tested comes from an underlying normal population. H_A — the underlying population is non-normal (note that the form of this alternative population is not specified).

Underlying assumptions: All observations come from a single normal population.

When to use: To test normality on a set of raw measurements or following transformation of the data on the transformed scale. It can also be used on the residuals from a one-way ANOVA to test the joint normality of the groups being compared.

Steps involved: 1) Construct a normal probability plot of the dataset; 2) Calculate the correlation between the pairs on the probability plot; 3) Compare the test statistic against an α-level critical point; 4) Values higher than the critical point are consistent with the null hypothesis of normality, while values lower than the critical point suggest a non-normal fit.

Advantages/Disadvantages: Filliben's procedure is an excellent test of normality, with very similar characteristics to the Shapiro-Wilk test. As a correlation on a probability plot, the Filliben's test statistic will tend to be close to one (and more indicative of normality) when a probability plot of the same data exhibits a close-to-linear pattern. Critical points for Filliben's test are available for sample sizes up to 100. A table of special coefficients is not needed to run Filliben's test, only the ability to compute inverse normal probabilities.

SHAPIRO-WILK MULTIPLE GROUP TEST (SECTION 10.7)

Basic purpose: Diagnostic tool and a formal normality goodness-of-fit test for multiple groups.

Hypothesis tested: H_0 — datasets being tested all come from underlying normal populations, possibly with different means and/or variances. H_A — at least one underlying population is non-normal (note that the form of this alternative population is not specified).

Underlying assumptions: The observations in each group all come from, possibly different, normal populations.

When to use: Can be used to test normality on multiple sets of raw measurements or, by first making a transformation, to test normality of the data groups on the transformed scale. It is particularly helpful when used in conjunction with Welch's *t*-test.

Steps involved: 1) Compute Shapiro-Wilk statistic (**Section 10.5**) on each group separately; 2) Transform the Shapiro-Wilk statistics into *z*-scores and combine into an omnibus *z*-score; 3) Compare the test statistic against an α-level critical point; 4) Values higher than the critical point are consistent with the null hypothesis of normality for all the populations, while values lower than the critical point suggest a non-normal fit of one or more groups.

Advantages/Disadvantages: As an extension of the Shapiro-Wilk test, the multiple group test shares many of its desirable properties. Users should be careful, however, not to assume that a result consistent with the hypothesis of normality implies that all groups follow the ***same*** normal distribution. The multiple group test does not assume that all groups have the same means or variances. Special coefficients are needed to convert Shapiro-Wilk statistics into *z*-scores, but once converted, no other special tables needed to run test besides a standard normal table.

LEVENE'S TEST (SECTION 11.2)

Basic purpose: Diagnostic tool. Levene's test is a formal numerical test of equality of variances across multiple populations.

Hypothesis tested: H_0 — The population variances across all the datasets being tested are equal. H_A — One or more pairs of population variances are unequal.

Underlying assumptions: The data set from each population is assumed to be roughly normal in distribution. Since Levene's test is designed to work well even with somewhat non-normal data (***i.e.***, it is fairly robust to non-normality), precise normality is not an overriding concern.

When to use: Levene's method can be used to test the equal variance assumption underlying one-way ANOVA for a group of wells. Used in this way, the test is run on the absolute values of the residuals after first subtracting the mean of each group being compared. If Levene's test is significant, the original data may need to be transformed to stabilize the variances before running an ANOVA.

Steps involved: 1) Compute the residuals of each group by subtracting the group mean; 2) conduct a one-way ANOVA on the absolute values of the residuals; and 3) if the ANOVA *F*-statistic is significant at the 5% α-level, conclude the underlying population variances are unequal. If not, conclude the data are consistent with the null hypothesis of equal variances.

Advantages/Disadvantages: As a test of equal variances, Levene's test is reasonably robust to non-normality. It is much more so than for Bartlett's test (recommended within the 1989 ***Interim Final***

Guidance [IFG]). In addition, Levene's method uses the same basic equations as those needed to run a one-way ANOVA.

MEAN-STANDARD DEVIATION SCATTER PLOT (SECTION 11.3)

Basic purpose: Diagnostic tool. It is a graphical method to examine degree of association between mean levels and standard deviations at a series of wells. Positive correlation or association between these quantities is known as a 'proportional effect' and is characteristic of skewed distributions such as the lognormal.

Hypothesis tested: Though not a formal test, the mean-standard deviation scatter plot provides a visual indication of whether variances are roughly equal from well to well, or whether the variance depends on the well mean.

Underlying Assumptions: None.

When to use: Useful as a graphical indication of 1) equal variances or 2) proportional effects between the standard deviation and mean levels. A positive correlation between well means and standard deviations may signify that a transformation is needed to stabilize the variances.

Steps involved: 1) Compute the sample mean and standard deviation for each well; 2) plot the mean-standard deviation pairs on a scatter plot; and 3) examine the plot for any association between the two quantities.

Advantages/Disadvantages: Not a formal test of homoscedasticity (*i.e.*, equal variances). It is helpful in assessing whether a transformation might be warranted to stabilize unequal variances.

DIXON'S TEST (SECTION 12.3)

Basic purpose: Diagnostic tool. It is used to identify (single) outliers within smaller datasets.

Hypothesis tested: H_0 — Outlier(s) comes from same normal distribution as rest of the dataset. H_A — Outlier(s) comes from different distribution than rest of the dataset.

Underlying assumptions: Data without the suspected outlier(s) are normally distributed. Test recommended only for sample sizes up to 25.

When to use: Try Dixon's test when one value in a dataset appears anomalously low or anomalously high when compared to the other data values. Be cautious about screening apparent high outliers in compliance point wells. Even if found to be statistical outliers, such extreme concentrations may represent contamination events. A safer application of outlier tests is with background or baseline samples. Even then, always try to establish a physical reason for the outlier if possible (*e.g.*, analytical error, transcription mistake, *etc.*).

Steps involved: 1) Remove the suspected outlier and test remaining data for normality. If non-normal, try a transformation to achieve normality; 2) Once remaining data are normal, calculate Dixon's statistic, depending on the sample size n; 3) Compare Dixon's statistic against an α-level critical point; and 4) If Dixon's statistic exceeds the critical point, conclude the suspected value is a statistical outlier. Investigate this measurement further.

Advantages/Disadvantages: Dixon's test is only recommended for sample sizes up to 25. Furthermore, if there is more than one outlier, Dixon's test may lead to masking (*i.e.*, a non-significant result) where two or more outliers close in value 'hide' one another. If more than one outlier is suspected, always test the *least* extreme value first.

ROSNER'S TEST (SECTION 12.4)

Basic purpose: Diagnostic tool. It is used to identify multiple outliers within larger datasets.

Hypothesis tested: H_0 — Outliers come from same normal distribution as the rest of the dataset. H_A — Outliers come from different distribution than the rest of the dataset.

Underlying assumptions: Data without the suspected outliers are normally distributed. Test recommended for sample sizes of at least 20.

When to use: Try Rosners's test when multiple values in a dataset appear anomalously low or anomalously high when compared to the other data values. As Dixon's test, be cautious about screening apparent high outliers in compliance point wells. Always try to establish a physical reason for an outlier if possible (*e.g.*, analytical error, transcription mistake, *etc.*).

Steps involved: 1) Identify the maximum number of possible outliers ($r_0 \leq 5$) and the number of suspected outliers ($r \leq r_0$). Remove the suspected outliers and test the remaining data for normality. If non-normal, try a transformation to achieve normality; 2) Once remaining data are normal, successively compute the mean and standard deviation, removing the next most extreme value each time until r_0 possible outliers have been removed; 3) Compute Rosner's statistic based on the number (r) of suspected outliers; and 4) If Rosner's statistic exceeds an α-level critical point, conclude there are r statistical outliers. Investigate these measurements further. If Rosner's statistic does not exceed the critical point, recompute the test for (r–1) possible outliers, successively reducing r until either the critical point is exceeded or $r = 0$.

Advantages/Disadvantages: Rosner's test is only recommended for sample sizes of 20 or more, but can be used to identify up to 5 outliers per use. It is more complicated to use than some other outlier tests, but does not require special tables other than to determine α-level critical points.

ONE-WAY ANALYSIS OF VARIANCE [ANOVA] FOR SPATIAL VARIATION (SECTION 13.2.2)

Basic purpose: Diagnostic tool. Test to compare population means at multiple wells, in order to gauge the presence of spatial variability.

Hypothesis tested: H_0 — Population means across all tested wells are equal. H_A — One or more pairs of population means are unequal.

Underlying assumptions: 1) ANOVA residuals at each well or group must be normally distributed using the original data or after transformation. Residuals should be tested for normality using a goodness-of-fit procedure; 2) population variances across all wells must be equal. This assumption can be tested with box plots and Levene's test; and 3) each tested well should have at least 3 to 4 separate observations.

When to use: The one-way ANOVA procedure can be used to identify significant spatial variation across a group of distinct well locations. The method is particularly useful for a group of multiple upgradient wells, to determine whether or not there are large average concentration differences from one location to the next due to natural groundwater fluctuations and/or differences in geochemistry. If downgradient wells are included in an ANOVA, the downgradient groundwater should not be contaminated, at least if a test of *natural* spatial variation is desired. Otherwise, a significant difference in population means could reflect the presence of either recent or historical contamination.

Steps involved: 1) Form the ANOVA residuals by subtracting from each measurement its sample well mean; 2) test the ANOVA residuals for normality and equal variance. If either of these assumptions is violated, try a transformation of the data and retest the assumptions; 3) compute the one-way ANOVA *F*-statistic; 4) if the *F*-statistic exceeds an α-level critical point, conclude the null hypothesis of equal population means has been violated and that there is some (perhaps substantial) degree of spatial variation; 5) if the *F*-statistic does not exceed the critical point, conclude that the well averages are close enough to treat the combined data as coming from the same statistical population.

Advantages/Disadvantages: One-way ANOVA is an excellent technique for identifying differences in separate well populations, as long as the assumptions are generally met. However, a finding of significant spatial variability does not specify the reason for the well-to-well differences. Additional information or investigation may be necessary to determine why the spatial differences exist. Be especially careful when (1) testing a combination of upgradient and downgradient wells that downgradient contamination is not the source of the difference found with ANOVA; and 2) when ANOVA identifies significant spatial variation and intrawell tests are called for. In the latter case, the ANOVA results can sometimes be used to estimate more powerful intrawell prediction and control limits. Such an adjustment comes directly from the ANOVA computations, requiring no additional calculation.

ANALYSIS OF VARIANCE [ANOVA] FOR TEMPORAL EFFECTS (SECTIONS 14.2.2 & 14.3.3)

Basic purpose: Diagnostic tool. It is a test to compare population means at multiple sampling events, after pooling the event data across wells. The test can also used to adjust data across multiple wells for common temporal dependence.

Hypothesis tested: H_0 — Population means across all sampling events are equal. H_A — One or more pairs of population means are unequal.

Underlying assumptions: 1) ANOVA residuals from the population at each sampling event must be normal or normalized. These should be tested for normality using a goodness-of-fit procedure; 2) the population variances across all sampling events must be equal. Test this assumption with box plots and Levene's test; and 3) each tested well should have at least 3 to 4 observations per sampling event.

When to use: 1) The ANOVA procedure for temporal effects should be used to identify significant temporal variation over a series of distinct sampling events. The method assumes that spatial variation by well location is not a significant factor (this should have already been tested). ANOVA for temporal effects should be used when a time series plot of a group of wells exhibits roughly parallel traces over time, indicating a time-related phenomenon affecting all the wells in a similar

way on any given sampling event. If a significant temporal effect is found, the results of the ANOVA can be employed to adjust the standard deviation estimate and the degrees of freedom quantities needed for further upgradient-to-downgradient comparisons; 2) compliance wells can be included in ANOVA for temporal effects, since the temporal pattern is assumed to affect all the wells on-site, regardless of gradient; and 3) residuals from ANOVA for temporal effects can be used to create adjusted, temporally-stationary measurements in order to eliminate the temporal dependence.

Steps involved: 1) Compute the mean (across wells) from data collected on each separate sampling event; 2) form the ANOVA residuals by subtracting from each measurement its sampling event mean; 3) test the ANOVA residuals for normality and equal variance. If either of these assumptions is violated, try a transformation of the data and retest the assumptions; 4) compute the one-way ANOVA *F*-statistic; 5) if the *F*-statistic exceeds an α-level critical point, conclude the null hypothesis of equal population means has been violated and that there is some (perhaps substantial) degree of temporal dependence; 6) compute the degrees of freedom adjustment factor and the adjusted standard deviation for use in *interwell* comparisons; 7) if the *F*-statistic does not exceed the critical point, conclude that the sampling event averages are close enough to treat the combined data as if there were no temporal dependence; and use the residuals, if necessary, to create adjusted, temporally-stationary measurements, regardless of the significance of the *F*-test (**Section 14.3.3**).

Advantages/Disadvantages: 1) One-way ANOVA for temporal effects is a good technique for identifying time-related effects among a group of wells. The procedure should be employed when a strong temporal dependence is indicated by parallel traces in time series plots; 2) if there is both temporal dependence and strong spatial variability, the ANOVA for temporal effects may be non-significant due to the added spatial variation. A two-way ANOVA for temporal and spatial effects might be considered instead; and 3) even if the ANOVA is non-significant, the ANOVA residuals can still be used to adjust data for apparent temporal dependence.

SAMPLE AUTOCORRELATION FUNCTION (SECTION 14.2.3)

Basic purpose: Diagnostic tool. This is a parametric estimate and test of autocorrelation (*i.e.*, time-related dependence) in a data series from a single population.

Hypothesis tested: H_0 — Measurements from the population are independent of sampling events (*i.e.*, they are not influenced by the time when the data were collected). H_A — The distribution of measurements is impacted by the time of data collection.

Underlying assumptions: Data should be approximately normal, with few non-detects. Sampling events represented in the sample should be fairly regular and evenly spaced in time.

When to use: When testing a data series from a single population (*e.g.*, a single well), the sample autocorrelation function (also known as the correlogram) can determine whether there is a significant temporal dependence in the data.

Steps involved: 1) Form overlapping ordered pairs from the data series by pairing measurements 'lagged' by a certain number of sampling events (*e.g.*, all pairs with measurements spaced by $k = 2$ sampling events); 2) for each distinct lag (k), compute the sample autocorrelation; 3) plot the

autocorrelations from **Step 2** by lag (*k*) on a scatter plot; and 4) count any autocorrelation as significantly different from zero if its absolute magnitude exceeds $2/\sqrt{n}$, where *n* is the sample size.

Advantages/Disadvantages: 1) The sample autocorrelation function provides a graphical test of temporal dependence. It can be used not only to identify autocorrelation, but also as a planning tool for adjusting the sampling interval between events. The smallest lag (*k*) at which the autocorrelation is insignificantly different from zero is the minimum sampling interval ensuring temporally uncorrelated data; 2) the test only applies to a single population at a time and cannot be used to identify temporal effects that span across groups of wells simultaneously. In that scenario, use a one-way ANOVA for temporal effects; and 3) tests for significant autocorrelation depend on the data being approximately normal; use the rank von Neumann ratio for non-normal samples.

RANK VON NEUMANN RATIO (SECTION 9.4)

Basic purpose: Diagnostic tool. It is a non-parametric test of first-order autocorrelation (*i.e.*, time-related dependence) in a data series from a single population.

Hypothesis tested: H_0 — Measurements from the population are independent of sampling events (*i.e.*, they are not influenced by the time when the data were collected). H_A — The distribution of measurements is impacted by the time of data collection.

Underlying assumptions: Data need not be normally distributed. However, it is assumed that the data series can be uniquely ranked according to concentration level. Ties in the data (*e.g.*, non-detects) are not technically allowed. Although a mid-rank procedure (as used in the Wilcoxon rank-sum test) to rank tied values might be considered, the available critical points for the rank von Neumann ratio statistic only directly apply to cases where a unique ranking is possible.

When to use: When testing a data series from a single population (*e.g.*, a single well) for use in, perhaps, an intrawell prediction limit, control chart, or test of trend, the rank von Neumann ratio can determine whether there is a significant temporal dependence in the data. If the dependence is seasonal, the data may be adjusted using a seasonal correction (**Section 14.3.3**). If the dependence is a linear trend, remove the estimated trend and re-run the rank von Neumann ratio on the trend residuals before concluding there are additional time-related effects. Complex dependence may require consultation with a professional statistician.

Steps involved: 1) Rank the measurements by concentration level, but then list the ranks in the order the samples were collected; 2) using the ranks, compute the von Neumann ratio; 3) if the rank von Neumann ratio exceeds an α-level critical point, conclude the data exhibit no significant temporal correlation. Otherwise, conclude that a time-related pattern does exist. Check for seasonal cycles or linear trends using time series plots. Consult a professional statistician regarding possible statistical adjustments if the pattern is more complex.

Advantages/Disadvantages: The rank von Neumann ratio, as opposed to other common time series methods for determining autocorrelation, is a non-parametric test based on using the ranks of the data instead of the actual concentration measurements. The test is simple to compute and can be used as a formal confirmation of temporal dependence, even if the autocorrelation appears fairly obvious on a time series plot. As a limiting feature, the test only applies to a single population at a time and

cannot be used to identify temporal effects that span across groups of wells simultaneously. In that scenario, a one-way ANOVA for temporal effects is a better diagnostic tool. Because critical points for the rank von Neumann ratio have not been developed for the presence of ties, the test will not be useful for datasets with substantial portions of non-detects.

DARCY EQUATION (SECTION 14.3.2)

Basic purpose: Method to determine a sampling interval ensuring that distinct physical volumes of groundwater are sampled on any pair of consecutive events.

Hypothesis tested: Not a statistical test or formal procedure.

Underlying assumptions: Flow regime is one in which Darcy's equation is approximately valid.

When to use: Use Darcy's equation to gauge the minimum travel time necessary for distinct volumes of groundwater to pass through each well screen. Physical independence of samples does not guarantee statistical independence, but it increases the likelihood of statistical independence. Use to design or plan for a site-specific sampling frequency, as well as what formal statistical tests and retesting strategies are possible given the amount of temporally-independent data that can be collected each evaluation period.

Steps involved: 1) Using knowledge of the site hydrogeology, calculate the horizontal and vertical components of average groundwater velocity with Darcy's equation; 2) Determine the minimum travel time needed between field samples to ensure physical independence; 3) Specify a sampling interval during monitoring no less than the travel time obtained via the Darcy computation.

Advantages/Disadvantages: Darcy's equation is relatively straightforward, but is not a statistical procedure. It is not applicable to certain hydrologic environments. Further, it is not a substitute for a direct estimate of autocorrelation. Statistical independence is not assured using Darcy's equation, so caution is advised.

SEASONAL CORRECTION (SECTION 14.3.3)

Basic purpose: Method to adjust a longer data series from a single population for an obvious seasonal cycle or fluctuation pattern. By removing the seasonal pattern, the remaining residuals can be used in further statistical procedures (*e.g.*, prediction limits, control charts) and treated as independent of the seasonal correlation.

Hypothesis tested: The seasonal correction is not a formal statistical test. Rather, it is a statistical adjustment to data for which a definite seasonal pattern has been identified.

Underlying assumptions: There should be enough data so that at least 3 full seasonal cycles are displayed on a time series plot. It is also assumed that the seasonal component has a stationary (*i.e.*, stable) mean and variance during the period of data collection.

When to use: Use the seasonal correction when a longer series of data must be examined, but a time series plot indicates a clearly recurring, seasonal fluctuation of concentration levels. If not removed, the seasonal dependence will tend to upwardly bias the estimated variability and could lead to inaccurate or insufficiently powerful tests.

Steps involved: 1) Using a time series plot of the data series, separate the values into common sampling events for each year (*e.g.*, all January measurements, all third quarter values, *etc.*); 2) compute the average of each subgroup and the overall mean of the dataset; and 3) adjust the data by removing the seasonal pattern.

Advantages/Disadvantages: The seasonal correction described in the Unified Guidance is relatively simple to perform and offers a more accurate standard deviation estimates compared to using unadjusted data. Removal of the seasonal component may reveal other previously unnoticed features of the data, such as a slow-moving trend. A fairly long data series is required to confirm the presence of a recurring seasonal cycle. Furthermore, many complex time-related patterns cannot be handled by this simple correction. In such cases, consultation with a professional statistician may be necessary.

SEASONAL MANN-KENDALL TEST FOR TREND (SECTION 14.3.4)

Basic purpose: Method for detection monitoring. It is used to identify the presence of a significant (upward) trend at a compliance point when data also exhibit seasonal fluctuations. It may also be used in compliance/assessment and corrective action monitoring to track upward or downward trends.

Hypothesis tested: H_0 — No discernible linear trend exists in the concentration data over time. H_A — A non-zero, (upward) linear component to the trend does exist.

Underlying assumptions: Since the seasonal Mann-Kendall trend test is a non-parametric method, the underlying data need not be normal or follow a particular distribution. No special adjustment for ties is needed.

When to use: Use when 1) upgradient-to-downgradient comparisons are inappropriate so that intrawell tests are called for; 2) a control chart or intrawell prediction limit cannot be used because of possible trends in the intrawell background, and 3) the data also exhibit seasonality. A trend test can be particularly helpful at sites with recent or historical contamination where it is uncertain if background is already contaminated. An upward trend in these cases will document the changing concentration levels more accurately than either a control chart or intrawell prediction limit, both of which assume a stationary background mean concentration.

Steps involved: 1) Divide the data into separate groups representing common sampling events from each year; 2) compute the Mann-Kendall test statistic (S) and its standard deviation ($SD[S]$) on each group; 3) sum the separate Mann-Kendall statistics into an overall test statistic; 4) compare this statistic against an α-level critical point; and 5) if the statistic exceeds the critical point, conclude that a significant upward trend exists. If not, conclude there is insufficient evidence for identifying a significant, non-zero trend.

Advantages/Disadvantages: 1) The seasonal Mann-Kendall test does not require any special treatment for non-detects, only that all non-detects be set to a common value lower than any of the detected values; and 2) the test is easy to compute and reasonably efficient for detecting (upward) trends in the presence of seasonality. Approximate critical points are derived from the standard normal distribution.

SIMPLE SUBSTITUTION (SECTION 15.2)

Basic purpose: A simple adjustment for non-detects in a dataset. One-half the reporting limit [RL] is substituted for each non-detect to provide a numerical approximation to the unknown true concentration.

Hypothesis tested: None.

Underlying assumptions: The true non-detect concentration is assumed to lie somewhere between zero and the reporting limit. Furthermore, that the probability of the true concentration being less than half the RL is about the same as the probability of it being greater than half the RL.

When to use: In general, simple substitution should be used when the dataset contains a relatively small proportion of non-detects, say no more than 10-15%. Use with larger non-detect proportions can result in biased estimates, especially if most of the detected concentrations are recorded at low levels (*e.g.*, at or near RL).

Steps involved: 1) Determine the reporting limit; and 2) replace each non-detect with one-half RL as a numerical approximation.

Advantages/Disadvantages: Simple substitution of half the RL is the easiest adjustment available for non-detect data. However, it can lead to biased estimates of the mean and particularly the variance if employed when more than 10-15% of the data are non-detects.

CENSORED PROBABILITY PLOT (SECTIONS 15.3 AND 15.4)

Basic purpose: Diagnostic tool. It is a graphical fit to normality of a mixture of detected and non-detect measurements. Adjustments are made to the plotting positions of the detected data under the assumption that all measurements come from a common distributional model.

Hypothesis tested: As a graphical tool, the censored probability plot is not a formal statistical test. However, it can provide an indication as to whether a dataset is consistent with the hypothesis that the mixture of detects and non-detects come from the same distribution and that the non-detects make up the lower tail of that distribution.

Underlying assumptions: Dataset consists of a mixture of detects and non-detects, all arising from a common distribution. Data must be normal or normalized.

When to use: Use the censored probability plot to check the viability of the Kaplan-Meier or robust regression on order statistics [ROS] adjustments for non-detect measurements. If the plot is linear, the data are consistent with a model in which the unobserved non-detect concentrations comprise the lower tail of the underlying distribution.

Steps involved: 1) Using either Kaplan-Meier or ROS, construct a partial ranking of the detected values to account for the presence of non-detects; 2) determine standard normal quantiles that match the ranking of the detects; and 3) graph the detected values against their matched normal quantiles on a probability plot and examine for a linear fit.

Advantages/Disadvantages: The censored probability plot offers a visual indication of whether a mixture of detects and non-detects come from the same (normal) distribution. There are, however, no formal critical points to aid in deciding when the fit is 'linear enough.' Correlation coefficients can be computed to informally aid the assessment. Censored probability plots can also be constructed on transformed data to help select a normalizing transformation.

KAPLAN-MEIER ADJUSTMENT (SECTION 15.3)

Basic purpose: Diagnostic tool. It is used to adjust a mixture of detected and non-detect data for the unknown concentrations of non-detect values. The Kaplan-Meier procedure leads to adjusted estimates for the mean and standard deviation of the underlying population.

Hypothesis tested: As a statistical adjustment procedure, the Kaplan-Meier method is not a formal statistical test. Rather, it allows estimation of characteristics of the population by assuming the combined group of detects and non-detects come from a common distribution.

Underlying assumptions: Dataset consists of a mixture of detects and non-detects, all arising from the same distribution. Data must be normal or normalized in the context of the Unified Guidance. Kaplan-Meier should not be used when more than 50% of the data are non-detects.

When to use: Since the Kaplan-Meier adjustment assumes all the measurements arise from the same statistical process, but that some of these measurements (*i.e.*, the non-detects) are unobservable due to limitations in analytical technology, Kaplan-Meier should be used when this model is the most realistic or reasonable choice. In particular, when constructing prediction limits, confidence limits, or control charts, the mean and standard deviation of the underlying population must be estimated. If non-detects occur in the dataset (but do not account for more than half of the observations), the Kaplan-Meier adjustment can be used to determine these estimates, which in turn can be utilized in constructing the desired statistical test.

Steps involved: 1) Sort the detected values and compute the 'risk set' associated with each detect; 2) using the risk set, compute the Kaplan-Meier cumulative distribution function [CDF] estimate associated with each detect; 3) calculate adjusted estimates of the population mean and standard deviation using the Kaplan-Meier CDF values; and 4) use these adjusted population estimates in place of the sample mean and standard deviation in prediction limits, confidence limits, and control charts.

Advantages/Disadvantages: Kaplan-Meier offers a way to adjust for significant fractions of non-detects without having to know the actual non-detect concentration values. It is more difficult to use than simple substitution, but avoids the biases inherent in that method.

ROBUST REGRESSION ON ORDER STATISTICS [ROS] (SECTION 15.4)

Basic purpose: Diagnostic tool. It is a method to adjust mixture of detects and non-detects for the unknown concentrations of non-detect values. Robust ROS leads to adjusted estimates for the mean and standard deviation of the underlying population by imputing a distinct estimated value for each non-detect.

Hypothesis tested: As a statistical adjustment procedure, robust ROS is not a formal statistical test. Rather, it allows estimation of characteristics of the population by assuming the combined group of detects and non-detects come from a common distribution.

Underlying assumptions: Dataset consists of a mixture of detects and non-detects, all arising from the same distribution. Data must be normal or normalized in the context of the Unified Guidance. Robust ROS should not be used when more than 50% of the data are non-detects.

When to use: Since robust regression on order statistics assumes all the measurements arise from the same statistical process, robust ROS should be used when this model is reasonable. In particular, when constructing prediction limits, confidence limits, or control charts, the mean and standard deviation of the underlying population must be estimated. If non-detects occur in the dataset (but do not account for more than half of the observations), robust ROS can be used to determine these estimates, which in turn can be utilized to construct the desired statistical test.

Steps involved: 1) Sort the distinct reporting limits [RL] for non-detect values and compute 'exceedance probabilities' associated with each RL; 2) using the exceedance probabilities, compute 'plotting positions' for the non-detects, essentially representing CDF estimates associated with each RL; 3) impute values for individual non-detects based on their RLs and plotting positions; 4) compute adjusted mean and standard deviation estimates via the sample mean and standard deviation of the combined set of detects and imputed non-detects; and 5) use these adjusted population estimates in place of the (unadjusted) sample mean and standard deviation in prediction limits, confidence limits, and control charts.

Advantages/Disadvantages: Robust ROS offers an alternative to Kaplan-Meier to adjust for significant fractions of non-detects without having to know the actual non-detect concentration values. It is more difficult to use than simple substitution, but avoids the biases inherent in that method.

COHEN'S METHOD AND PARAMETRIC ROS (SECTION 15.5)

Basic purpose: Diagnostic tools. These are other methods to adjust mixture of detects and non-detects to obtain the unknown mean and standard deviation for the entire data set

Hypothesis tested: **N**either technique is a formal statistical test. Rather, they allow estimation of characteristics of the population by assuming the combined group of detects and non-detects come from a common distribution.

Underlying assumptions: Dataset consists of a mixture of detects and non-detects, all arising from the same distribution. Data must be normal or normalized in the context of the Unified Guidance. Neither should be used when more than 50% of the data are non-detects nor when data contain multiple non-detect levels.

When to use: Since these methods assume that all the measurements arise from the same statistical process, they should be used when this model is reasonable. In particular, when constructing prediction limits, confidence limits, or control charts, the mean and standard deviation of the underlying population must be estimated. If non-detects occur in the dataset (but do not account for more than half of the observations), they can be used to determine these estimates, which in turn can be utilized to construct the desired statistical test.

Steps involved: Cohen's Method: 1) data are sorted into non-detect and detected portions; 2) detect mean and standard deviation estimates are calculated; 3) intermediate quantities of the ND% and a factor γ are calculated and used to locate the appropriate λ value from a table; and 4) full data set mean and standard deviation estimates are then obtained using formulas based on the detected mean, standard deviation, the detection limit and λ. Parametric ROS: 1) detected data are sorted in ascending order; 2) standardized normal distribution Z-values are generated from the full set of ranked values. Those corresponding to the sorted detected values are retained; 3) the detected values are then regressed against the Z-values; and 4) the resulting regression intercept and slope are the estimates of the mean and standard deviation for the full data set.

Advantages/Disadvantages: These two methods offer alternatives to Kaplan-Meier and robust ROS. The key limitation is that only data containing a single censoring limit can be used. In some situations using logarithmic data, their application can lead to biased estimates of the mean and standard deviation. Where appropriate, these methods are less computationally intensive that either Kaplan-Meier or robust ROS.

POOLED VARIANCE T-TEST (SECTION 16.1.1)

Basic purpose: Method for detection monitoring. This test compares the means of two populations.

Hypothesis tested: H_0 — Means of the two populations are equal; H_A — Means of the two populations are unequal (for the usual one-sided alternative, the hypothesis would state that the mean of the second population is greater than the mean of the first population).

Underlying assumptions: 1) The data from each population must be normal or normalized; 2) when used for interwell tests, there should be no significant spatial variability; 3) at least 4 observations per well should be available before applying the test; and 4) the two group variances are equal.

When to use: The pooled variance *t*-test can be used to test for groundwater contamination at very small sites, those consisting of maybe 3 or 4 wells and monitoring for 1 or 2 constituents. Site configurations with larger combinations of wells and constituents should employ a retesting scheme using either prediction limits or control charts. The pooled variance *t*-test can also be used to test proposed updates to intrawell background. A ***non-significant*** *t*-test in this latter case suggests the two sets of data are sufficiently similar to allow the initial background to be updated by augmenting with more recent measurements.

Steps involved: 1) Test the combined residuals from each population for normality. Make a data transformation if necessary; 2) test for equal variances, and if equal, compute a pooled variance estimate; 3) compute the pooled variance *t*-statistic and the degrees of freedom; 3) compare the *t*-statistic against a critical point based on both the α-level and the degrees of freedom; and 4) if the *t*-statistic exceeds the critical point, conclude the null hypothesis of equal means has been violated.

Advantages/Disadvantages: 1) The pooled variance *t*-test is one of the easiest to compute *t*-test procedures, but requires an assumption of equal variances across both populations; 2) because the *t*-test is a well-understood statistical procedure, the Unified Guidance recommends its use at very small groundwater monitoring facilities. For larger sites, however, repeated use of the *t*-test at a given α-level will lead to an unacceptably high risk of false positive error; and 3) if substantial spatial variability exists, the use of any *t*-test for upgradient-to-downgradient comparisons may lead

to inaccurate conclusions. A significant difference in the population averages could also indicate the presence of natural geochemical factors differentially affecting the concentration levels at different wells. In these situations, consider an intrawell test instead.

WELCH'S T-TEST (SECTION 16.1.2)

Basic purpose: Method for detection monitoring. This test compares the means of two populations.

Hypothesis tested: H_0 — Means of the two populations are equal; H_A — Means of the two populations are unequal (for the usual one-sided alternative, the hypothesis would state that the mean of the second population is greater than the mean of the first population).

Underlying assumptions: 1) The data from each population must be normal or normalized; 2) when used for interwell tests, there should be no significant spatial variability; and 3) At least 4 observations per well should be available before applying the test.

When to use: Welch's *t*-test can be used to test for groundwater contamination at very small sites, those consisting of maybe 3 or 4 wells and monitoring for 1 or 2 constituents. Site configurations with larger combinations of wells and constituents should employ a retesting scheme using either prediction limits or control charts. Welch's *t*-test can also be used to test proposed updates to intrawell background data. A ***non-significant*** *t*-test in this latter case suggests the two sets of data are sufficiently similar to allow the initial background to be updated by augmenting with the more recent measurements.

Steps involved: 1) Test the combined residuals from each population for normality. Make a data transformation if necessary; 2) compute Welch's *t*-statistic and approximate degrees of freedom; 3) compare the *t*-statistic against a critical point based on both the α-level and the estimated degrees of freedom; and 4) if the *t*-statistic exceeds the critical point, conclude the null hypothesis of equal means has been violated.

Advantages/Disadvantages: 1) Welch's *t*-test is slightly more difficult to compute than other common *t*-test procedures, but has the advantage of ***not*** requiring equal variances across both populations. Furthermore, it has been shown to perform statistically as well or better than other *t*-tests; 2) it can be used at very small groundwater monitoring facilities, but should be avoided at larger sites. Repeated use of the *t*-test at a given α-level will lead to an unacceptably high risk of false positive error; and 3) if there is substantial spatial variability, use of Welch's *t*-test for interwell tests may lead to inaccurate conclusions. A significant difference in the population averages may reflect the presence of natural geochemical factors differentially affecting the concentration levels at different wells. In these situations, consider an intrawell test instead.

WILCOXON RANK-SUM TEST (SECTION 16.2)

Basic purpose: Method for detection monitoring. This test compares the medians of two populations.

Hypothesis tested: H_0 — Both populations have equal medians (and, in fact, are identical in distribution). H_A — The two population medians are unequal (in the usual one-sided alternative, the hypothesis would state that the median of the second population is larger than the median of the first).

Underlying assumptions: 1) While the Wilcoxon rank-sum test does not require normal data, it does assume both populations have the same distributional form and that the variances are equal. If the data are non-normal but there at most a few non-detects, the equal variance assumption may be tested through the use of box plots and/or Levene's test. If non-detects make-up a large fraction of the observations, equal variances may have to be assumed rather than formally verified; 2) use of the Wilcoxon rank-sum procedure for interwell tests assumes there is no significant spatial variability. This is more likely to be the case in precisely those circumstances where the Wilcoxon procedure might be used: when there are high fractions of non-detects, so that most of the concentration measurements at any location are at low levels; and 3) there should be at least 4 background measurements and at least 2-4 compliance point values.

When to use: The Wilcoxon rank-sum test can be used to test for groundwater contamination at very small sites, those consisting of maybe 3 or 4 wells and monitoring for 1 or 2 constituents. Site configurations with larger combinations of wells and constituents should employ a retesting scheme using non-parametric prediction limits. Note, however, that non-parametric prediction limits often require large background sample sizes to be effective. The Wilcoxon rank-sum can be useful when a high percentage of the data is non-detect, but the amount of available background data is limited. Indeed, an *intrawell* Wilcoxon procedure may be helpful in some situations where the false positive rate would otherwise be too high to run intrawell prediction limits.

Steps involved: 1) Rank the combined set of values from the two datasets, breaking ties if necessary by using midranks; 2) compute the sum of the ranks from the compliance point well and calculate the Wilcoxon test statistic; 3) compare the Wilcoxon test statistic against an α-level critical point; and 4) if the test statistic exceeds the critical point, conclude that the null hypothesis of equal medians has been violated.

Advantages/Disadvantages: 1) The Wilcoxon rank-sum test is an excellent technique for small sites with constituent non-detect data. Compared to other possible methods such as the test of proportions or exact binomial prediction limits, the Wilcoxon rank-sum does a better job overall of correctly identifying elevated groundwater concentrations while limiting false positive error; 2) because the Wilcoxon rank-sum is easy to compute and understand, the Unified Guidance recommends its use at very small groundwater monitoring facilities. For larger sites, repeated use of the Wilcoxon rank-sum at a given α-level will lead to an unacceptably high risk of false positive error; and 3) if substantial spatial variability exists, the use of the Wilcoxon rank-sum for interwell tests may lead to inaccurate conclusions. A significant difference in the population medians may signal the presence of natural geochemical differences rather than contaminated groundwater. In these situations, consider an intrawell test instead.

TARONE-WARE TEST (SECTION 16.3)

Basic purpose: Non-parametric method for detection monitoring. This is an extension of Wilcoxon rank-sum, an alternative test to compare the medians in two populations when non-detects are prevalent.

Hypothesis tested: H_0 — Both populations have equal medians (and, in fact, are identical in distribution). H_A — The two population medians are unequal (in the usual one-sided alternative, the hypothesis would state that the median of the second population is larger than the median of the first).

Underlying assumptions: 1) The Tarone-Ware test does not require normal data, but does assume both populations have the same distributional form and that the variances are equal; and 2) use of the Tarone-Ware procedure for interwell tests assumes there is no significant spatial variability. This is more likely to be the case when there are high fractions of data non-detects, so that most of the concentration measurements at any location are at low and similar levels.

When to use: The Tarone-Ware test can be used to test for groundwater contamination at very small sites, those consisting of perhaps 3 or 4 wells and monitoring for 1 or 2 constituents. Site configurations with larger combinations of wells and constituents should employ a retesting scheme using non-parametric prediction limits. Note, however, that non-parametric prediction limits often require large background sample sizes to be effective. The Tarone-Ware test can be useful when a high percentage of the data is non-detect, but the amount of available background data is limited. The Tarone-Ware test is also an alternative to the Wilcoxon rank-sum when there are multiple reporting limits and/or it is unclear how to fully rank the data as required by the Wilcoxon.

Steps involved: 1) Sort the distinct detected values in the combined data set; 2) count the 'risk set' associated with each distinct value from **Step 1** and compute the expected number of compliance point detections within each risk set; 3) form the Tarone-Ware test statistic from the expected counts in **Step 2**; 4) compare the test statistic against a standard normal α-level critical point; and 5) if the test statistic exceeds the critical point, conclude that the null hypothesis of equal medians has been violated.

Advantages/Disadvantages: The Tarone-Ware test is an excellent technique for small sites with constituent non-detect data having multiple reporting limits. If substantial spatial variability exists, use of the Tarone-Ware test for interwell tests may lead to inaccurate conclusions. A significant difference in the population medians may signal the presence of natural geochemical differences rather than contaminated groundwater. In these situations, consider an intrawell test instead.

ONE-WAY ANALYSIS OF VARIANCE [ANOVA] (SECTION 17.1.1)

Basic purpose: Formal interwell detection monitoring test and diagnostic tool. It compares population means at multiple wells, in order to detect contaminated groundwater when tested against background.

Hypothesis tested: H_0 — Population means across all tested wells are equal. H_A — One or more pairs of population means are unequal.

Underlying assumptions: 1) ANOVA residuals at each well or population must be normally distributed or transformable to normality. These should be tested for normality using a goodness-of-fit procedure; 2) the population variances across all wells must be equal. This assumption can be tested with box plots and Levene's test; and 3) each tested well should have at least 3 to 4 separate observations.

When to use: The one-way ANOVA can sometimes be used to identify to simultaneously test for contaminated groundwater across a group of distinct well locations. As an inherently interwell test, ANOVA should be utilized only on constituents exhibiting little to no spatial variation. Most uses of ANOVA have been superseded by prediction limits and control charts, although it is commonly employed to identify spatial variability or temporal dependence across a group of wells.

Steps involved: 1) Form the ANOVA residuals by subtracting from each measurement its sample well mean; 2) tcst the ANOVA residuals for normality and equal variance. If either of these assumptions is violated, try a transformation of the data and retest the assumptions; 3) compute the one-way ANOVA *F*-statistic; 4) if the *F*-statistic exceeds an α-level critical point, conclude the null hypothesis of equal population means has been violated and that at least one pair of wells shows a significant difference in concentration levels; and 5) test each compliance well individually to determine which one or more exceeds background.

Advantages/Disadvantages: ANOVA is only likely to be infrequently used to make upgradient-to-downgradient comparisons in formal detection monitoring testing. The regulatory restrictions for per-constituent α-levels using ANOVA make it difficult to adequately control site-wide false positive rates [SWFPR]. Even if spatial variability is not a significant problem, users are advised to consider interwell prediction limits or control charts, and to incorporate some form of retesting

KRUSKAL-WALLIS TEST (SECTION 17.1.2)

Basic purpose: Formal interwell detection monitoring test and diagnostic tool. It compares population medians at multiple wells, in order to detect contaminated groundwater when tested against background. It is also useful as a non-parametric alternative to ANOVA for identifying spatial variability in constituents with non-detects or for data that cannot be normalized.

Hypothesis tested: H_0 — Population medians across all tested wells are equal. H_A — One or more pairs of population medians are unequal.

Underlying assumptions: 1) As a non-parametric alternative to ANOVA, data need not be normal; 2) the population variances across all wells must be equal. This assumption can be tested with box plots and Levene's test if the non-detect proportion is not too high; and 3) each tested well should have at least 3 to 4 separate observations.

When to use: The Kruskal-Wallis test can sometimes be used to identify to simultaneously test for contaminated groundwater across a group of distinct well locations. As an inherently interwell test, Kruskal-Wallis should be utilized for this purpose only with constituents exhibiting little to no spatial variation. Most uses of the Kruskal-Wallis (similar to ANOVA) have been superseded by prediction limits, although it can be used to identify spatial variability and/or temporal dependence across a group of wells when the sample data are non-normal or have higher proportions of non-detects.

Steps involved: 1) Sort and form the ranks of the combined measurements; 2) compute the rank-based Kruskal-Wallis test statistic (*H*); 3) if the *H*-statistic exceeds an α-level critical point, conclude the null hypothesis of equal population medians has been violated and that at least one pair of wells shows a significant difference in concentration levels; and 5) test each compliance well individually to determine which one or more exceeds background.

Advantages/Disadvantages: 1) The Kruskal-Wallis test is only likely to be infrequently used to make upgradient-to-downgradient comparisons in formal detection monitoring testing. The regulatory restrictions for per-constituent α-levels using ANOVA make it difficult to adequately control the SWFPR. Even if spatial variability is not a significant problem, users are advised to consider

interwell prediction limits, and to incorporate some form of retesting; and 2) the Kruskal-Wallis test can be used to test for spatial variability in constituents with significant fractions of non-detects.

TOLERANCE LIMIT (SECTION 17.2.1)

Basic purpose: Formal interwell detection monitoring test of background versus one or more compliance wells. Tolerance limits can be used as an alternative to one-way ANOVA. These can also be used in corrective action as an alternative clean-up limit.

Hypothesis tested: H_0 — Population means across all tested wells are equal. H_A — One or more pairs of population means are unequal.

Underlying assumptions: 1) Data should be normal or normalized; 2) the population variances across all wells are assumed to be equal. This assumption can be difficult to test when comparing a single new observation from each compliance well against a tolerance limit based on background; and 3) there should be a minimum of 4 background measurements, preferably 8-10 or more.

When to use: A tolerance limit can be used in place of ANOVA for detecting contaminated groundwater. It is more flexible than ANOVA since 1) as few as one new measurement per compliance well is needed to run a tolerance limit test, and 2) no post-hoc testing is necessary to identify which compliance wells are elevated over background. Most uses of tolerance limits (similar to ANOVA) have been superseded by prediction limits, due to difficulty of incorporating retesting into tolerance limit schemes. If a hazardous constituent requires a background-type standard in compliance/assessment or corrective action, a tolerance limit can be computed on background and used as a fixed GWPS.

Steps involved: 1) Compute background sample mean and standard deviation; 2) calculate upper tolerance limit on background with high confidence and high coverage; 3) collect one or more observations from each compliance well and test each against the tolerance limit; and 4) identify a well as contaminated if any of its observations exceed the tolerance limit.

Advantages/Disadvantages: Tolerance limits are likely to be used only infrequently to be used as either interwell or intrawell tests. Prediction limits or control charts offer better control of false positive rates, and less is known about the impact of retesting on tolerance limit performance.

NON-PARAMETRIC TOLERANCE LIMIT (SECTION 17.2.2)

Basic purpose: Formal interwell detection monitoring test of background versus one or more compliance wells. Non-parametric tolerance limits can be used as an alternative to the Kruskal-Wallis test. They may also be used in compliance/assessment or corrective action to define a background GWPS.

Hypothesis tested: H_0 — Population medians across all tested wells are equal. H_A — One or more pairs of population medians are unequal.

Underlying assumptions: 1) As a non-parametric test, non-normal data with non-detects can be used; and 2) there should be a minimum of 8-10 background measurements and preferably more.

When to use: A non-parametric tolerance limit can be used in place of the Kruskal-Wallis test for detecting contaminated groundwater. It is more flexible than Kruskal-Wallis since 1) as few as one new measurement per compliance well is needed to run a tolerance limit test, and 2) no post-hoc testing is necessary to identify which compliance wells are elevated over background. Most uses of tolerance limits have been superseded by prediction limits, due to difficulty of incorporating retesting into tolerance limit schemes. However, when a clean-up limit cannot or has not been specified in corrective action, a tolerance limit can be computed on background and used as a site-specific alternate concentration limit [ACL].

Steps involved: 1) Compute a large order statistic from background and set this value as the upper tolerance limit; 2) calculate the confidence and coverage associated with the tolerance limit; 3) collect one or more observations from each compliance well and test each against the tolerance limit; and 4) identify a well as contaminated if any of its observations exceed the tolerance limit.

Advantages/Disadvantages: 1) Tolerance limits are likely to be used only infrequently to be used as either interwell or intrawell tests. Prediction limits or control charts offer better control of false positive rates, and less is known about the impact of retesting on tolerance limit performance; and 2) non-parametric tolerance limits have the added disadvantage of generally requiring large background samples to ensure adequate confidence and/or coverage. For this reason, it is strongly recommended that a parametric tolerance limit be constructed whenever possible.

LINEAR REGRESSION (SECTION 14.4)

Basic purpose: Method for detection monitoring and diagnostic tool. It is used to identify the presence of a significantly increasing trend at a compliance point or any trend in background data sets.

Hypothesis tested: H_0 — No discernible linear trend exists in the concentration data over time. H_A — A non-zero, (upward) linear component to the trend does exist.

Underlying assumptions: Trend residuals should be normal or normalized, equal in variance, and statistically independent. If a small fraction of non-detects exists (≤10-15%), use simple substitution to replace each non-detect by half the reporting limit [RL]. Test homoscedasticity of residuals with a *scatter plot* (**Section 9.1**).

When to use: Use a test for trend when 1) upgradient-to-downgradient comparisons are inappropriate so that intrawell tests are called for, and 2) a control chart or intrawell prediction limit cannot be used because of possible trends in the intrawell background. A trend test can be particularly helpful at sites with recent or historical contamination where it is uncertain to what degree intrawell background is already contaminated. The presence of an upward trend in these cases will document the changing nature of the concentration data much more accurately than either a control chart or intrawell prediction limit, both of which assume a stable baseline concentration.

Steps involved: 1) If a linear trend is evident on a time series plot, construct the linear regression equation; 2) subtract the estimated trend line from each observation to form residuals; 3) test residuals for assumptions listed above; and 4) test regression slope to determine whether it is significantly different from zero. If so and the slope is positive, conclude there is evidence of a significant upward trend.

Advantages/Disadvantages: Linear regression is a standard statistical method for identifying trends and other linear associations between pairs of random variables. However, it requires approximate normality of the trend residuals. Confidence bands around regression trends can be used in compliance/assessment and corrective action to determine compliance with fixed standards even when concentration levels are actively changing (*i.e.*, when a trend is apparent).

MANN-KENDALL TEST FOR TREND (SECTION 17.3.2)

Basic purpose: Method for detection monitoring and diagnostic tool. It is used to identify the presence of a significant (upward) trend at a compliance point or any trend in background data.

Hypothesis tested: H_0 — No discernible linear trend exists in the concentration data over time. H_A — A non-zero, (upward) linear component to the trend does exist.

Underlying assumptions: Since the Mann-Kendall trend test is a non-parametric method, the underlying data need not be normal or follow any particular distribution. No special adjustment for ties is needed.

When to use: Use a test for trend when 1) interwell tests are inappropriate so that intrawell tests are called for, and 2) a control chart or intrawell prediction limit cannot be used because of possible trends in intrawell background. A trend test can be particularly helpful at sites with recent or historical contamination where it is uncertain if intrawell background is already contaminated. An upward trend in these cases documents changing concentration levels more accurately than either a control chart or intrawell prediction limit, both of which assume a stationary background mean concentration.

Steps involved: 1) Sort the data values by time of sampling/collection; 2) consider all possible pairs of measurements from different sampling events; 3) score each pair depending on whether the later data point is higher or lower in concentration than the earlier one, and sum the scores to get Mann-Kendall statistic; 4) compare this statistic against an α-level critical point; and 5) if the statistic exceeds the critical point, conclude that a significant upward trend exists. If not, conclude there is insufficient evidence for identifying a significant, non-zero trend.

Advantages/Disadvantages: The Mann-Kendall test does not require any special treatment for non-detects, only that all non-detects can be set to a common value lower than any of the detects. The test is easy to compute and reasonably efficient for detecting (upward) trends. Exact critical points are provided in the Unified Guidance for $n \leq 20$; a normal approximation can be used for $n > 20$. 3) A version of the Mann-Kendall test (the seasonal Mann-Kendall, **Section 14.3.4**) can be used to test for trends in data that exhibit seasonality.

THEIL-SEN TREND LINE (SECTION 17.3.3)

Basic purpose: Method for detection monitoring. This is a non-parametric alternative to linear regression for estimating a linear trend.

Hypothesis tested: As presented in the Unified Guidance, the Theil-Sen trend line is not a formal hypothesis test but rather an estimation procedure. The algorithm can be modified to formally test whether the true slope is significantly different from zero, but this question will already be answered if used in conjunction with the Mann-Kendall procedure.

Underlying assumptions: Like the Mann-Kendall trend test, the Theil-Sen trend line is non-parametric, so the underlying data need not be normal or follow a particular distribution. Furthermore, data ranks are not used, so no special adjustment for ties is needed.

When to use: It is particularly helpful when used in conjunction with the Mann-Kendall test for trend. The latter test offers information about whether a trend exists, but does not estimate the trend line itself. Once a trend is identified, the Theil-Sen procedure indicates how quickly the concentration level is changing with time.

Steps involved: 1) Sort the data set by date/time of sampling; 2) for each pair of distinct sampling events, compute the simple pairwise slope; 3) sort the list of pairwise slopes and set the overall slope estimate (Q) as the median slope in this list; 4) compute the median concentration and the median date/time of sampling; and 5) construct the Theil-Sen trend as the line passing through the median scatter point from **Step 4** with slope Q.

Advantages/Disadvantages: Although non-parametric, the Theil-Sen slope estimator does not use data ranks but rather the concentrations themselves. The method is non-parametric because the median pairwise slope is utilized, thus ignoring extreme values that might otherwise skew the slope estimate. The Theil-Sen trend line is as easy to compute as the Mann-Kendall test and does not require any special adjustment for ties (***e.g.***, non-detects).

PREDICTION LIMIT FOR *M* FUTURE VALUES (SECTION 18.2.1)

Basic purpose: Method for detection monitoring. This technique estimates numerical bound(s) on a series of m independent future values. The prediction limit(s) can be used to test whether the mean of one or more compliance well populations are equal to the mean of a background population.

Hypothesis tested: H_0 — The true mean of m future observations arises from the same population as the mean of measurements used to construct the prediction limit. H_A — The m future observations come from a distribution with a different mean than the population of measurements. Since an upper prediction limit is of interest in detection monitoring, the alternative hypothesis would state that the future observations are distributed with a larger mean than the background population.

Underlying assumptions: 1) Data used to construct the prediction limit must be normal or normalized. Adjustments for small to moderate fractions of non-detects can be made, perhaps using Kaplan-Meier or robust ROS; 2) although the variances of both populations (background and future values) are assumed to be equal, rarely will there be enough data from the future population to verify this assumption except during periodic updates to background; and 3) if used for upgradient-to-downgradient comparisons, there should be no significant spatial variability.

When to use: Prediction limits on individual observations can be used as an alternative in detection monitoring to either one-way ANOVA or Dunnett's multiple comparison with control [MCC] procedure. Assuming there is insignificant natural spatial variability, an interwell prediction limit can be constructed using upgradient or other representative background data. The number of future samples (m) should be chosen to reflect a single new observation collected from each downgradient or compliance well prior to the next statistical evaluation, plus a fixed number (m–1) of possible resamples. The initial future observation at each compliance point is then compared against the prediction limit. If it exceeds the prediction limit, one or more resamples are collected from the

'triggered' well and also tested against the prediction limit. If substantial spatial variability exists, prediction limits for individual values can be constructed on a well-specific basis using intrawell background. The larger the intrawell background size, the better. To incorporate retesting, it must be feasible to collect up to (m–1) additional, but independent, resamples from each well.

Steps involved: 1) Compute the estimated mean and standard deviation of the background data; 2) considering the type of prediction limit (*i.e.*, interwell or intrawell), the number of future samples m, the desired site-wide false positive rate, and the number of wells and monitoring parameters, determine the prediction limit multiplier (κ); 3) compute the prediction limit as the background mean plus κ times the background standard deviation; and 4) compare each initial future observation against the prediction limit. If both the initial measurement and resample(s) exceed the limit, conclude the null hypothesis of equal means has been violated.

Advantages/Disadvantages: Prediction limits for individual values offer several advantages compared to the traditional one-way ANOVA and Dunnett's multiple comparison with control [MCC] procedures. Prediction limits are not bound to a minimum 5% per-constituent false positive rate and can be constructed to meet a target site-wide false positive rate [SWFPR] while maintaining acceptable statistical power. Unlike the one-way ANOVA F-test, only the comparisons of interest (*i.e.*, each compliance point against background) are tested. This gives the prediction limit more statistical power. Prediction limits can be designed for intrawell as well as interwell comparisons.

PREDICTION LIMIT FOR FUTURE MEAN (SECTION 18.2.2)

Basic purpose: Method for detection monitoring or compliance monitoring. It is used to estimate numerical limit(s) on an independent mean constructed from p future values. The prediction limits(s) can be used to test whether the mean of one population is equal to the mean of a separate (background) population.

Hypothesis tested: H_0 — The true mean of p future observations arise from the same population as the mean of measurements used to construct the prediction limit. H_A — The p future observations come from a distribution with a different mean than the population of background measurements. Since an upper prediction limit is of interest in both detection and compliance monitoring, the alternative hypothesis would state that the future observations are distributed with a larger mean than that of the background population.

Underlying assumptions: 1) Data used to construct the prediction limit must be normal or normalized. Adjustments for small to moderate fractions of non-detects can be made, perhaps using Kaplan-Meier or robust ROS; 2) although the variances of both populations (background and future values) are assumed to be equal, rarely will there be enough data from the future population to verify this assumption; and 3) if used for upgradient-to-downgradient comparisons, there should be no significant spatial variability.

When to use: Prediction limits on means can be used as an alternative in detection monitoring to either one-way ANOVA or Dunnett's multiple comparison with control [MCC] procedure. Assuming there is insignificant natural spatial variability, an interwell prediction limit can be constructed using upgradient or other representative background data. The number of future samples p should be chosen to reflect the number of samples that will be collected at each compliance well prior to the next statistical evaluation (*e.g.*, 2, 4, *etc.*). The average of these p observations at each compliance

point is then compared against the prediction limit. If it is feasible to collect at least *p* additional, but independent, resamples from each well, retesting can be incorporated into the procedure by using independent mean(s) of *p* samples as confirmation value(s).

If substantial spatial variability exists, prediction limits for means can be constructed on a well-specific basis using intrawell background. At least two future values must be available per well. Larger intrawell background size are preferable. To incorporate retesting, it must be feasible to collect at least *p* independent resamples from each well, in addition to the initial set of *p* samples. A prediction limit can also be used in some compliance monitoring settings when a fixed compliance health based limit cannot be use and the compliance point data must be compared directly to a background GWPS. In this case, the compliance point mean concentration is tested against an upper prediction limit computed from background. No retesting would be employed for this latter kind of test.

Steps involved: 1) Compute the background sample mean and standard deviation; 2) considering the type of prediction limit (*i.e.*, interwell or intrawell), the number of future samples *p*, use of retesting, the desired site-wide false positive rate, and the number of wells and monitoring parameters, determine the prediction limit multiplier (κ); 3) compute the prediction limit as the background mean plus κ times the background standard deviation; 4) compare each future mean of order *p* (*i.e.*, a mean constructed from *p* values) against the prediction limit; and 5) if the future mean exceeds the limit and retesting is not feasible (or if used for compliance monitoring), conclude the null hypothesis of equal means has been violated. If retesting is feasible, conclude the null hypothesis has been violated only when the resampled mean(s) of order *p* also exceeds the prediction limit.

Advantages/Disadvantages: Prediction limits on means offer several advantages compared to the traditional one-way ANOVA and Dunnett's multiple comparison with control [MCC] procedure: Prediction limits are not bound to a minimum 5% per-constituent false positive rate. As such, prediction limits can be constructed to meet a target SWFPR, while maintaining acceptable statistical power. Unlike the one-way *F*-test, only the comparisons of interest (*i.e.*, each compliance point against background) are tested, giving the prediction limit more statistical power. Prediction limits can be designed for intrawell as well as interwell comparisons. One slight disadvantage is that ANOVA combines compliance point data with background to give a somewhat better per-well estimate of variability. But even this disadvantage can be overcome when using an interwell prediction limit by first running ANOVA on the combined background and compliance point data to generate a better variance estimate with a larger degree of freedom. A disadvantage compared to prediction limits on individual future values is that two or more new compliance point observations per well must be available to run the prediction limit on means. If only one new measurement per evaluation period can be collected, the user should instead construct a prediction limit on individual values.

NON-PARAMETRIC PREDICTION LIMIT FOR *M* FUTURE VALUES (SECTION 18.3.1)

Basic purpose: Method for detection monitoring. It is a non-parametric technique to estimate numerical limits(s) on a series of *m* independent future values. The prediction limit(s) can be used to test whether two samples are drawn from the same or different populations.

Hypothesis tested: H_0 — The *m* future observations come from the same distribution as the measurements used to construct the prediction limit. H_A — The *m* future observations come from a

different distribution than the population of measurements used to build the prediction limit. Since an upper prediction limit is of interest in detection monitoring, the alternative hypothesis is that the future observations are distributed with a larger median than the background population.

Underlying assumptions: 1) The data used to construct the prediction limit need not be normal; however, the forms of the both the background distribution and the future distribution are assumed to be the same. Since the non-parametric prediction limit is constructed as an order statistic of background, high fractions of non-detects are acceptable; 2) although the variances of both populations (background and future values) are assumed to be equal, rarely will there be enough data from the future population to verify this assumption; and 3) if used for upgradient-to-downgradient comparisons, there should be no significant spatial variability. Spatial variation is less likely to be significant in many cases where constituent data are primarily non-detect, allowing the use of a non-parametric interwell prediction limit test.

When to use: Prediction limits on individual values can be used as a non-parametric alternative in detection monitoring to either one-way ANOVA or Dunnett's multiple comparison with control [MCC] procedure. Assuming there is insignificant natural spatial variability, an interwell prediction limit can be constructed using upgradient or other representative background data. The number of future samples m should be chosen to reflect a single new observation collected from each compliance well prior to the next statistical evaluation, plus a fixed number (m–1) of possible resamples. The initial future observation at each compliance point is then compared against the prediction limit. If it exceeds the prediction limit, one or more resamples are collected from the 'triggered' well and also compared to the prediction limit.

Steps involved: 1) Determine the maximum, second-largest, or other highly ranked value in background and set the non-parametric prediction limit equal to this level; 2) considering the number of future samples m, and the number of wells and monitoring parameters, determine the achievable site-wide false positive rate [SWFPR]. If the error rate is not acceptable, consider possibly enlarging the pool of background data used to construct the limit or increasing the number of future samples m; 3) compare each initial future observation against the prediction limit; and 4) if both the initial measurement and resample(s) exceed the limit, conclude the null hypothesis of equal distributions has been violated.

Advantages/Disadvantages: Non-parametric prediction limits on individual values offer distinct advantages compared to the Kruskal-Wallis non-parametric ANOVA test. Prediction limits are not bound to a minimum 5% per-constituent false positive rate. As such, prediction limits can be constructed to meet a target SWFPR, while maintaining acceptable statistical power. Unlike the Kruskal-Wallis test, only the comparisons of interest (*i.e.*, each compliance point against background) are tested, giving the prediction limit more statistical power. Non-parametric prediction limits have the disadvantage of generally requiring fairly large background samples to effectively control false positive error and ensure adequate power.

PREDICTION LIMIT FOR FUTURE MEDIAN (SECTION 18.3.2)

Basic purpose: Method for detection monitoring and compliance monitoring. This is a non-parametric technique to estimate numerical limits(s) on the median of p independent future values. The prediction limit(s) is used to test whether the median of one or more compliance well populations is equal to the median of the background population.

Hypothesis tested: H_0 — The true median of p future observations arise from the same population as the median of measurements used to construct the prediction limit. H_A — The p future observations come from a distribution with a different median than the background population of measurements. Since an upper prediction limit is of interest in both detection monitoring and compliance monitoring, the alternative hypothesis is that the future observations are distributed with a larger median than the background population.

Underlying assumptions: 1) The data used to construct the prediction limit need not be normal; however, the forms of the both the background distribution and the future distribution are assumed to be the same. Since the non-parametric prediction limit is constructed as an order statistic of background, high fractions of non-detects are acceptable: 2) although the variances of both populations (background and future values) are assumed to be equal, rarely will there be enough data from the future population to verify this assumption; and 3) if used for upgradient-to-downgradient comparisons, there should be no significant spatial variability.

When to use: Prediction limits on medians can be used as a non-parametric alternative in detection monitoring to either one-way ANOVA or Dunnett's multiple comparison with control [MCC] procedure. Assuming there is insignificant natural spatial variability, an interwell prediction limit can be constructed using upgradient or other representative background data. The number of future samples p should be odd and chosen to reflect the number of samples that will be collected at each compliance well prior to the next statistical evaluation (*e.g.*, 3). The median of these p observations at each compliance point is then compared against the prediction limit. If it is feasible to collect at least p additional, but independent, resamples from each well, retesting can be incorporated into the procedure by using independent median(s) of p samples as confirmation value(s). A prediction limit for a compliance point median can also be constructed in certain compliance monitoring settings, when no fixed health-based compliance limit can be used and the compliance point data must be directly compared against a background GWPS. In this case, the compliance point median concentration is compared to an upper prediction limit computed from background. No retesting is employed for this latter kind of test.

Steps involved: 1) Determine the maximum, second-largest, or other highly ranked value in background and set the non-parametric prediction limit equal to this level; 2) considering the number of future samples p, whether or not retesting will be incorporated, and the number of wells and monitoring parameters, determine the achievable SWFPR. If the error rate is not acceptable, increase the background sample size or consider a non-parametric prediction limit on individual future values instead; 3) compare each future median of order p (*i.e.*, a median of p values) against the prediction limit; and 4) if the future median exceeds the limit and retesting is not feasible (or if the test is used for compliance monitoring), conclude the null hypothesis of equal medians has been violated. If retesting is feasible, conclude the null hypothesis has been violated only when the resampled median(s) of order p also exceeds the prediction limit.

Advantages/Disadvantages: Non-parametric prediction limits on medians offer distinct advantages compared to the Kruskal-Wallis test (a non-parametric one-way ANOVA). Prediction limits are not bound to a minimum 5% per-constituent false positive rate. As such, prediction limits can be constructed to meet a target SWFPR, while maintaining acceptable statistical power. Unlike the Kruskal-Wallis test, only the comparisons of interest (*i.e.*, each compliance point against background) are tested, giving the prediction limit more statistical power. A disadvantage in

detection monitoring compared to non-parametric prediction limits on individual future values is that at least three new compliance point observations per well must be available to run the prediction limit on medians. If only one new observation per evaluation period can be collected, construct instead a non-parametric prediction limit for individual values. All non-parametric prediction limits have the disadvantage of usually requiring fairly large background samples to effectively control false positive error and ensure adequate power.

SHEWHART-CUSUM CONTROL CHART (SECTION 20.2)

Basic purpose: Method for detection monitoring. These are used to quantitatively and visually track concentrations at a given well over time to determine whether they exceed a critical threshold (*i.e.*, control limit), thus implying a significant increase above background conditions.

Hypothesis tested: H_0 — Data plotted on the control chart follow the same distribution as the background data used to compute the baseline chart parameters. H_A — Data plotted on the chart follow a different distribution with higher mean level than the baseline data.

Underlying assumptions: Data used to construct the control chart must be approximately normal or normalized. Adjustments for small to moderate fractions of non-detects, perhaps using Kaplan-Meier or ROS, can be acceptable. There should be no discernible trend in the baseline data used to calculate the control limit.

When to use: Use control charts as an alternative to parametric prediction limits, when 1) there are enough uncontaminated baseline data to compute an accurate control limit, and 2) there are no trends in intrawell background. Retesting can be incorporated into control charts by judicious choice of control limit. This may need to be estimated using Monte Carlo simulations.

Steps involved: 1) Compute the intrawell baseline mean and standard deviation; 2) calculate an appropriate control limit from these baseline parameters, the desired retesting strategy and number of well-constituent pairs in the network; 3) construct the chart, plotting the control limit, the compliance point observations, and the cumulative sums [CUSUM]; and 4) determine that the null hypothesis is violated when either an individual concentration measurement or the cumulative sum exceeds the control limit.

Advantages/Disadvantages: Unlike prediction limits, control charts offer an explicit visual tracking of compliance point values over time and provide a method to judge whether these concentrations have exceeded a critical threshold. The Shewhart portion of the chart is especially good at detecting sudden concentration increases, while the CUSUM portion is preferred for detecting slower, steady increases over time. No non-parametric version of the combined Shewhart-CUSUM control chart exists, so non-parametric prediction limits should be considered if the data cannot be normalized.

CONFIDENCE INTERVAL AROUND NORMAL MEAN (SECTION 21.1.1)

Basic purpose: Method for compliance/assessment monitoring or corrective action. This is a technique for estimating a range of concentration values from sample data, in which the true mean of a normal population is expected to occur at a certain probability.

Hypothesis tested: In compliance monitoring, H_0 — True mean concentration at the compliance point is no greater than the predetermined groundwater protection standard [GWPS]. H_A — True mean

concentration is greater than the GWPS. In corrective action, H_0 — True mean concentration at the compliance point is greater than or equal to the fixed GWPS. H_A — True mean concentration is less than or equal to the fixed standard.

Underlying assumptions: 1) Compliance point data are approximately normal in distribution. Adjustments for small to moderate fractions of non-detects, perhaps using Kaplan-Meier or ROS, are encouraged; 2) data do not exhibit any significant trend over time; 3) there are a minimum of 4 observations for testing. Generally, at least 8 to 10 measurements are recommended; and 4) the fixed GWPS is assumed to represent a true mean average concentration, rather than a maximum or upper percentile.

When to use: A mean confidence interval can be used for normal data to determine whether there is statistically significant evidence that the average is either above a fixed GWPS (in compliance monitoring) or below the fixed standard (in corrective action). In either case, the null hypothesis is rejected only when the *entire* confidence interval lies on one or the other side of the GWPS. The key determinant in compliance monitoring is whether the *lower* confidence limit exceeds the GWPS, while in corrective action the *upper* confidence limit lies below the clean-up standard. Because of bias introduced by transformations when estimating a mean, this approach should not be used for highly-skewed or non-normal data. Instead consider a confidence interval around a lognormal mean or a non-parametric confidence interval. It is also not recommended for use when the data exhibit a significant trend. In that case, the estimate of variability will likely be too high, leading to an unnecessarily wide interval and possibly little chance of deciding the hypothesis. When a trend is present, consider instead a confidence interval around a trend line.

Steps involved: 1) Compute the sample mean and standard deviation; 2) based on the sample size and choice of a confidence level ($1-\alpha$), calculate either the lower confidence limit (for use in compliance monitoring) or the upper confidence limit (for use in corrective action); 3) compare the confidence limit against the GWPS or clean-up standard; and 4) if the lower confidence limit exceeds the GWPS in compliance monitoring or the upper confidence limit is below the clean-up standard, conclude that the null hypothesis should be rejected.

Advantages/Disadvantages: Use of a confidence interval instead of simply the sample mean for comparison to a fixed standard accounts for both the level of statistical variation in the data and the desired or targeted confidence level. The same basic test can be used both to document contamination above the compliance standard in compliance/assessment and to show a sufficient decrease in concentration levels below the clean-up standard in corrective action.

CONFIDENCE INTERVAL ON LOGNORMAL GEOMETRIC MEAN (SECTION 21.1.2)

Basic purpose: Method for compliance/assessment monitoring or corrective action. It is a technique to estimate the range of concentration values from sample data, in which the true geometric mean of a lognormal population is expected to occur at a certain probability.

Hypothesis tested: In compliance monitoring, H_0 — True mean concentration at the compliance point is no greater than the fixed compliance or groundwater protection standard [GWPS]. H_A — True mean concentration is greater than the GWPS. In corrective action, H_0 — True mean concentration at the compliance point is greater than the fixed compliance or clean-up standard. H_A — True mean concentration is less than or equal to the fixed standard.

Underlying assumptions: 1) Compliance point data are approximately lognormal in distribution. Adjustments for small to moderate fractions of non-detects, perhaps using Kaplan-Meier or ROS, are encouraged; 2) data do not exhibit any significant trend over time; 3) there are a minimum of 4 observations. Generally, at least 8 to 10 measurements are recommended; and 4) the fixed GWPS is assumed to represent a true geometric mean average concentration following a lognormal distribution, rather than a maximum or upper percentile. The GWPS also represents the true median.

When to use: A confidence interval on the geometric mean can be used for lognormal data to determine whether there is statistically significant evidence that the geometric average is either above a fixed numerical standard (in compliance monitoring) or below a fixed standard (in corrective action). In either case, the null hypothesis is rejected only when the *entire* confidence interval is to one side of the compliance or clean-up standard. Because of this fact, the key question in compliance monitoring is whether the *lower* confidence limit exceeds the GWPS, while in corrective action the user must determine whether the *upper* confidence limit is below the clean-up standard. Because of bias introduced by transformations when estimating the arithmetic lognormal mean, and the often unreasonably high upper confidence limits generated by Land's method for lognormal mean confidence intervals (see below), this approach is an alternative approach for lognormal data. One could also consider a non-parametric confidence interval. It is also not recommended for use when data exhibit a significant trend. In that case, the estimate of variability will likely be too high, leading to an unnecessarily wide interval and possibly little chance of deciding the hypothesis. When a trend is present, consider instead a confidence interval around a trend line.

Steps involved: 1) Compute the sample log-mean and log-standard deviation; 2) based on the sample size and choice of confidence level ($1-\alpha$), calculate either the lower confidence limit (for use in compliance monitoring) or the upper confidence limit (for use in corrective action) using the logged measurements and exponentiate the result; 3) compare the confidence limit against the GWPS or clean-up standard; and 4) if the lower confidence limit exceeds the GWPS in compliance monitoring or the upper confidence limit is below the clean-up standard, conclude that the null hypothesis should be rejected.

Advantages/Disadvantages: Use of a confidence interval instead of simply the sample geometric mean for comparison to a fixed standard accounts for both statistical variation in the data and the targeted confidence level. The same basic test can be used both to document contamination above the compliance standard in compliance/assessment and to show a sufficient decrease in concentration levels below the clean-up standard in corrective action.

CONFIDENCE INTERVAL ON LOGNORMAL ARITHMETIC MEAN (SECTION 21.1.3)

Basic purpose: Test for compliance/assessment monitoring or corrective action. This is a method by Land (1971) used to estimate the range of concentration values from sample data, in which the true arithmetic mean of a lognormal population is expected to occur at a certain probability.

Hypothesis tested: In compliance monitoring, H_0 — True mean concentration at the compliance point is no greater than the fixed compliance or groundwater protection standard [GWPS]. H_A — True mean concentration is greater than the GWPS. In corrective action, H_0 — True mean concentration at the compliance point is greater than the fixed compliance or clean-up standard. H_A — True mean concentration is less than or equal to the fixed standard.

Underlying assumptions: 1) Compliance point data are approximately lognormal in distribution. Adjustments for small to moderate fractions of non-detects, perhaps using Kaplan-Meier or ROS, are encouraged; 2) data do not exhibit any significant trend over time; 3) there are a minimum of 4 observations. Generally, at least 8 to 10 measurements are strongly recommended; and 4) the fixed GWPS is assumed to represent the true arithmetic mean average concentration, rather than a maximum or upper percentile.

When to use: Land's confidence interval procedure can be used for lognormally-distributed data to determine whether there is statistically significant evidence that the average is either above a fixed numerical standard (in compliance monitoring) or below a fixed standard (in corrective action). In either case, the null hypothesis is rejected only when the *entire* confidence interval is to one side of the compliance or clean-up standard. Because of this fact, the key question in compliance monitoring is whether the *lower* confidence limit exceeds the GWPS, while in corrective action the user must determine whether the *upper* confidence limit is below the clean-up standard. Because the lognormal distribution can have a highly skewed upper tail, this approach should only be used when the data fit the lognormal model rather closely, especially if used in corrective action. Consider instead a confidence interval around the lognormal geometric mean or a non-parametric confidence interval if this is not the case. It is also not recommended for data that exhibit a significant trend. In that situation, the estimate of variability will likely be too high, leading to an unnecessarily wide interval and possibly little chance of deciding the hypothesis. When a trend is present, consider instead a confidence interval around a trend line.

Steps involved: 1) Compute the sample log-mean and log-standard deviation; 2) based on the sample size, magnitude of the log-standard deviation and choice of confidence level ($1-\alpha$), determine Land's adjustment factor; 3) then calculate either the lower confidence limit (for use in compliance monitoring) or the upper confidence limit (for use in corrective action); 4) compare the confidence limit against the GWPS or clean-up standard; and 5) if the lower confidence limit exceeds the GWPS in compliance montoring or the upper confidence limit is below the clean-up standard, conclude that the null hypothesis should be rejected.

Advantages/Disadvantages: Use of a confidence interval instead of simply the sample mean for comparison to a fixed standard accounts for both statistical variation in the data and the targeted confidence level. The same basic test can be used both to document contamination above the compliance standard in compliance/assessment and to show a sufficient decrease in concentration levels below the clean-up standard in corrective action. Since the upper confidence limit on a lognormal mean can be extremely high for some populations, the user may need to consider a non-parametric upper confidence limit on the median concentration as an alternative or use a program such as **Pro-UCL** to determine an alternate upper confidence limit.

CONFIDENCE INTERVAL ON UPPER PERCENTILE (SECTION 21.1.4)

Basic purpose: Method for compliance monitoring. It is used to estimate the range of concentration values from sample data in which a pre-specified true proportion of a normal population is expected to occur at a certain probability. The test can also be used to identify the range of a true proportion or percentile (*e.g.*, the 95th) in population data which can be normalized.

Hypothesis tested: H_0 — True upper percentile concentration at the compliance point is no greater than the fixed compliance or groundwater protection standard [GWPS]. H_A — True upper percentile concentration is greater than the fixed GWPS.

Underlying assumptions: 1) Compliance point data are either normal in distribution or can be normalized. Adjustments for small to moderate fractions of non-detects, perhaps using Kaplan-Meier or ROS, are encouraged; 2) data do not exhibit any significant trend over time; 3) there are a minimum of at least 8 to 10 measurements; and 4) the fixed GWPS is assumed to represent a maximum or upper percentile, rather than an average concentration.

When to use: A confidence interval around an upper percentile can be used to determine whether there is statistically significant evidence that the percentile is above a fixed numerical standard. The null hypothesis is rejected only when the *entire* confidence interval is greater than the compliance standard. Because of this fact, the key question in compliance monitoring is whether the *lower* confidence limit exceeds the GWPS. This approach is not recommended for use when the data exhibit a significant trend. The estimate of variability will likely be too high, leading to an unnecessarily wide interval and possibly little chance of deciding the hypothesis.

Steps involved: 1) Compute the sample mean and standard deviation; 2) based on the sample size, pre-determined true proportion and test confidence level (1–α), calculate the lower confidence limit; 3) compare the confidence limit against the GWPS; and 4) if the lower confidence limit exceeds the GWPS, conclude that the true upper percentile is larger than the compliance standard.

Advantages/Disadvantages: If a fixed GWPS is intended to represent a 'not-to-be-exceeded' maximum or an upper percentile, statistical comparison requires the prior definition of a true or expected upper percentile against which sample data can be compared. Some standards may explicitly identify the expected percentile. The appropriate test then must estimate the confidence interval in which this true proportion is expected to lie. Either an upper or lower confidence limit can be generated, depending on whether compliance or corrective action hypothesis testing is appropriate. Whatever the interpretation of a given limit used as a GWPS, it should be determined in advance what a given standard represents before choosing which type of confidence interval to construct.

NON-PARAMETRIC CONFIDENCE INTERVAL ON MEDIAN (SECTION 21.2)

Basic purpose: Test for compliance/assessment monitoring or corrective action. It is a non-parametric method used to estimate the range of concentration values from sample data in which the true median of a population is expected to occur at a certain probability.

Hypothesis tested: In compliance monitoring, H_0 — True median concentration at the compliance point is no greater than the fixed compliance or groundwater protection standard [GWPS]. H_A — True median concentration is greater than the GWPS. In corrective action, H_0 — True median concentration at the compliance point is greater than the fixed compliance or clean-up standard. H_A — True median concentration is less than or equal to the fixed standard.

Underlying assumptions: 1) Compliance data need not be normal in distribution; up to 50% non-detects are acceptable; 2) data do not exhibit any significant trend over time; 3) there are a *minimum* of at least 7 measurements; and 4) the fixed GWPS is assumed to represent a true median average concentration, rather than a maximum or upper percentile.

When to use: A confidence interval on the median can be used for non-normal data (*e.g.*, samples with non-detects) to determine whether there is statistically significant evidence that the average (*i.e.*, median) is either above a fixed numerical standard (in compliance monitoring) or below a fixed standard (in corrective action). In either case, the null hypothesis is rejected only when the *entire* confidence interval is to one side of the compliance or clean-up standard. Because of this fact, the key question in compliance monitoring is whether the *lower* confidence limit exceeds the GWPS, while in corrective action the user must determine whether the *upper* confidence limit is below the clean-up standard. This approach is not recommended for use when data exhibit a significant trend. In that case, the variation in the data will likely be too high, leading to an unnecessarily wide interval and possibly little chance of deciding the hypothesis. It is also possible that the apparent trend is an artifact of differing detection or reporting limits that have changed over time. The trend may disappear if all non-detects are imputed at a common value or RL. If a trend is still present after investigating this possibility, but a significant portion of the data are non-detect, consultation with a professional statistician is recommended.

Steps involved: 1) Order and rank the data values; 2) pick tentative interval endpoints close to the estimated median concentration; 3) using the selected endpoints, compute the achieved confidence level of the lower confidence limit for use in compliance monitoring or that of the upper confidence limit for corrective action; 4) iteratively expand the interval until either the selected endpoints achieve the targeted confidence level or the maximum or minimum data value is chosen as the confidence limit; and 5) compare the confidence limit against the GWPS or clean-up standard. If the lower confidence limit exceeds the GWPS in compliance monitoring or the upper confidence limit is below the clean-up standard, conclude that the null hypothesis should be rejected.

Advantages/Disadvantages: Use of a confidence interval instead of simply the sample median for comparison to a fixed limit accounts for both statistical variation in the data and the targeted confidence level. The same basic test can be used both to document contamination above the compliance standard in compliance/assessment and to show a sufficient decrease in concentration levels below the clean-up standard in corrective action. By not requiring normal or normalized data, the non-parametric confidence interval can accommodate a substantial fraction of non-detects. A minor disadvantage is that a non-parametric confidence interval estimates the location of the median, instead of the mean. For symmetric populations, these quantities will be the same, but for skewed distributions they will differ. So if the compliance or clean-up standard is designed to represent a mean concentration, the non-parametric interval around the median may not provide a completely fair and/or accurate comparison. In some cases, the non-parametric confidence limit will not achieve the desired confidence level even if set to the maximum or minimum data value, leading to a higher risk of false positive error.

NON-PARAMETRIC CONFIDENCE INTERVAL ON UPPER PERCENTILE (SECTION 21.2)

Basic purpose: Non-parametric method for compliance monitoring. It is used to estimate the range of concentration values from sample data in which a pre-specified true proportion of a population is expected to occur at a certain probability. Exact probabilities will depend upon sample data ranks.

Hypothesis tested: H_0 — True upper percentile concentration at the compliance point is no greater than the fixed compliance or groundwater protection standard [GWPS]. H_A — True upper percentile concentration is greater than the GWPS.

Underlying assumptions: 1) Compliance point data need not be normal; large fractions of non-detects can be acceptable; 2) data do not exhibit any significant trend over time; 3) there are a minimum of at least 8 to 10 measurements; and 4) the fixed GWPS is assumed to represent a true upper percentile of the population, rather than an average concentration.

When to use: A confidence interval on an upper percentile can be used to determine whether there is statistically significant evidence that the percentile is above a fixed numerical standard. The null hypothesis is rejected only when the *entire* confidence interval is greater than the compliance standard. Because of this fact, the key determinant in compliance/assessment monitoring is whether the *lower* confidence limit exceeds the GWPS. This approach is not recommended for use when data exhibit a significant trend. In that case, the estimate of variability will likely be too high, leading to an unnecessarily wide interval and possibly little chance of deciding the hypothesis.

Steps involved: 1) Order and rank the data values; 2) select tentative interval endpoints close to the estimated upper percentile concentration; 3) using the selected endpoints, compute the achieved confidence level of the lower confidence limit; 4) iteratively expand the interval until either the selected lower endpoint achieves the targeted confidence level or the minimum data value is chosen as the confidence limit; and 5) compare the confidence limit against the GWPS. If the lower confidence limit exceeds the GWPS, conclude that the population upper percentile is larger than the compliance standard.

Advantages/Disadvantages: If a fixed GWPS is intended to represent a 'not-to-be-exceeded' maximum or an upper percentile, statistical comparison requires the prior definition of a true or expected upper percentile against which sample data can be compared. Some standards may explicitly identify the expected percentile. The appropriate test then must estimate the confidence interval in which this true proportion is expected to lie. Either an upper or lower confidence limit can be generated, depending on whether compliance or corrective action hypothesis testing is appropriate. Whatever the interpretation of a given limit used as a GWPS, it should be determined in advance what a given standard represents before choosing which type of confidence interval to construct. However, precise non-parametric estimation of upper percentiles often requires much larger sample sizes than the parametric option (**Section 21.1.4**). For this reason, a *parametric* confidence interval for upper percentile tests is recommended whenever possible, especially if a suitable transformation can be found or adjustments made for non-detect values.

CONFIDENCE BAND AROUND LINEAR REGRESSION (SECTION 21.3.1)

Basic purpose: Method for compliance/assessment monitoring or corrective action when stationarity cannot be assumed. It is used to estimate ranges of concentration values from sample data around each point of a predicted linear regression line at a specified probability. The prediction line (based on regression of concentration values against time) represents the best estimate of gradually changing true mean levels over the time period.

Hypothesis tested: In compliance monitoring, H_0 — True mean concentration at the compliance point is no greater than the fixed compliance or groundwater protection standard [GWPS]. H_A — True mean concentration is greater than the GWPS. In corrective action, H_0 — True mean concentration at the compliance point is greater than the fixed compliance or clean-up standard. H_A — True mean concentration is less than or equal to the fixed standard.

Underlying assumptions: 1) Compliance point values exhibit a linear trend with time, with normally distributed residuals. Use simple substitution with small (≤10-15%) fractions of non-detects. Non-detect adjustment methods are not recommended; 2) there are a minimum of 4 observations. Generally, at least 8 to 10 measurements are recommended; and 3) the fixed GWPS is assumed to represent an average concentration, rather than a maximum or upper percentile.

When to use: A confidence interval around a trend line should be used in cases where a linear trend is apparent on a time series plot of the compliance point data. Even if observed well concentrations are either increasing under compliance monitoring or decreasing in corrective action, it does not necessarily imply that the true mean concentration at the current time is either above or below the fixed GWPS. While the trend line properly accounts for the fact that the mean is changing with time, the null hypothesis is rejected only when the *entire* confidence interval is to one side of the compliance or clean-up standard at the most recent point(s) in time. The key determinant in compliance monitoring is whether the *lower* confidence limit at a specified point in time exceeds the GWPS, while in corrective action the *upper* confidence limit at a specific time must lie below the clean-up standard to be considered in compliance.

Steps involved: 1) Check for presence of a trend on a time series plot; 2) estimate the coefficients of the best-fitting linear regression line; 3) compute the trend line residuals and check for normality; 4) if data are non-normal, try re-computing the regression and residuals after transforming the data; 5) compute the lower confidence limit band around the trend line for compliance monitoring or the upper confidence limit band around the trend line for corrective action; and 6) compare the confidence limit at each sampling event against the GWPS or clean-up standard. If the lower confidence limit exceeds the GWPS in compliance/assessment or the upper confidence limit is below the clean-up standard on one or more recent sampling events, conclude that the null hypothesis should be rejected.

Advantages/Disadvantages: Use of a confidence interval around the trend line instead of simply the regression line itself for comparison to a fixed standard accounts for both statistical variation in the data and the targeted confidence level. The same basic test can be used both to document contamination above the compliance standard in compliance/assessment and to show a sufficient decrease in concentration levels below the clean-up standard in corrective action. By estimating the trend line first and then using the residuals to construct the confidence interval, variation due to the trend itself is removed, providing a more powerful test (via a narrower interval) of whether or not the true mean is on one side of the fixed standard. This technique can only be used when the identified trend is reasonably linear and the trend residuals are approximately normal.

NON-PARAMETRIC CONFIDENCE BAND AROUND THEIL-SEN TREND (SECTION 21.3.1)

Basic purpose: Non-parametric method for compliance/assessment or corrective action when stationarity cannot be assumed. It is used to estimate ranges of concentration values from sample data around each point of a predicted Theil-Sen trend line at a specified probability. The prediction line represents the best estimate of gradually changing true median levels over the time period.

Hypothesis tested: In compliance monitoring, H_0 — True mean concentration at the compliance point is no greater than the fixed compliance or groundwater protection standard [GWPS]. H_A — True mean concentration is greater than the GWPS. In corrective action, H_0 — True mean concentration at the

compliance point is greater than the fixed compliance or clean-up standard. H_A — True mean concentration is less than or equal to the fixed standard.

Underlying assumptions: 1) Compliance point values exhibit a linear trend with time; 2) non-normal data and substantial levels of non-detects up to 50% are acceptable; 3) there are a minimum of 8-10 observations available to construct the confidence band; and 4) the fixed GWPS is assumed to represent a median average concentration, rather than a maximum or upper percentile.

When to use: A confidence interval around a trend line should be used in cases where a linear trend is apparent on a time series plot of the compliance point data. Even if observed well concentrations are either increasing under compliance monitoring or decreasing in corrective action, it does not necessarily imply that the true mean concentration at the current time is either above or below the fixed GWPS. While the trend line properly accounts for the fact that the mean is changing with time, the null hypothesis is rejected only when the *entire* confidence interval is to one side of the compliance or clean-up standard at the most recent point(s) in time. The key determinant in compliance monitoring is whether the *lower* confidence limit at a specified point in time exceeds the GWPS, while in corrective action the *upper* confidence limit at a specific time must lie below the clean-up standard to be considered in compliance.

Steps involved: 1) Check for presence of a trend on a time series plot; 2) construct a Theil-Sen trend line; 3) use bootstrapping to create a large number of simulated Theil-Sen trends on the sample data; 4) construct a confidence band by selecting lower and upper percentiles from the set of bootstrapped Theil-Sen trend estimates; and 5) compare the confidence band at each sampling event against the GWPS or clean-up standard. If the lower confidence band exceeds the GWPS in compliance/assessment or the upper confidence band is below the clean-up standard on one or more recent sampling events, conclude that the null hypothesis should be rejected.

Advantages/Disadvantages: Use of a confidence band around the trend line instead of simply the Theil-Sen trend line itself for comparison to a fixed standard accounts for both statistical variation in the data and the targeted confidence level. The same basic test can be used both in compliance/assessment and in corrective action. By estimating the trend line first and then using bootstrapping to construct the confidence band, variation due to the trend itself is removed, providing a more powerful test (via a narrower interval) of whether or not the true mean is on one side of the fixed standard. This technique can only be used when the identified trend is reasonably linear. The Theil-Sen trend estimates the change in median level rather than the mean. For roughly symmetric populations, this will make little difference; for highly skewed populations, the trend in the median may not accurately reflect changes in mean concentration levels.