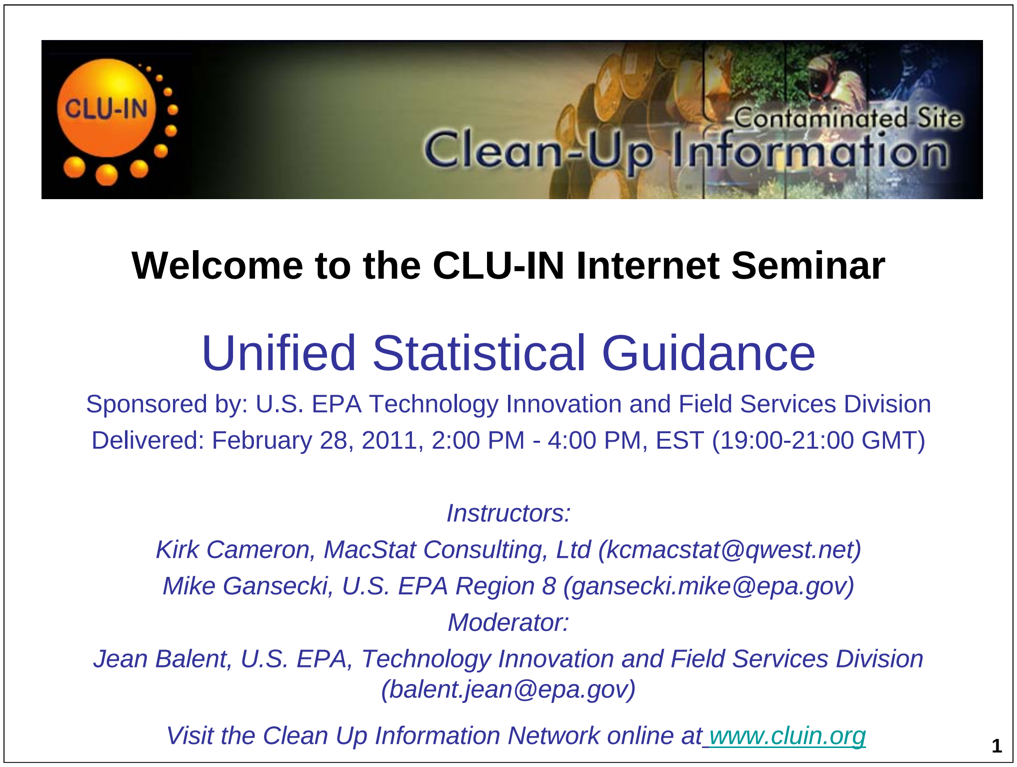

CLU-IN
Contaminated Site
Clean-Up Information

Welcome to the CLU-IN Internet Seminar

Unified Statistical Guidance

Sponsored by: U.S. EPA Technology Innovation and Field Services Division

Delivered: February 28, 2011, 2:00 PM - 4:00 PM, EST (19:00-21:00 GMT)

Instructors:

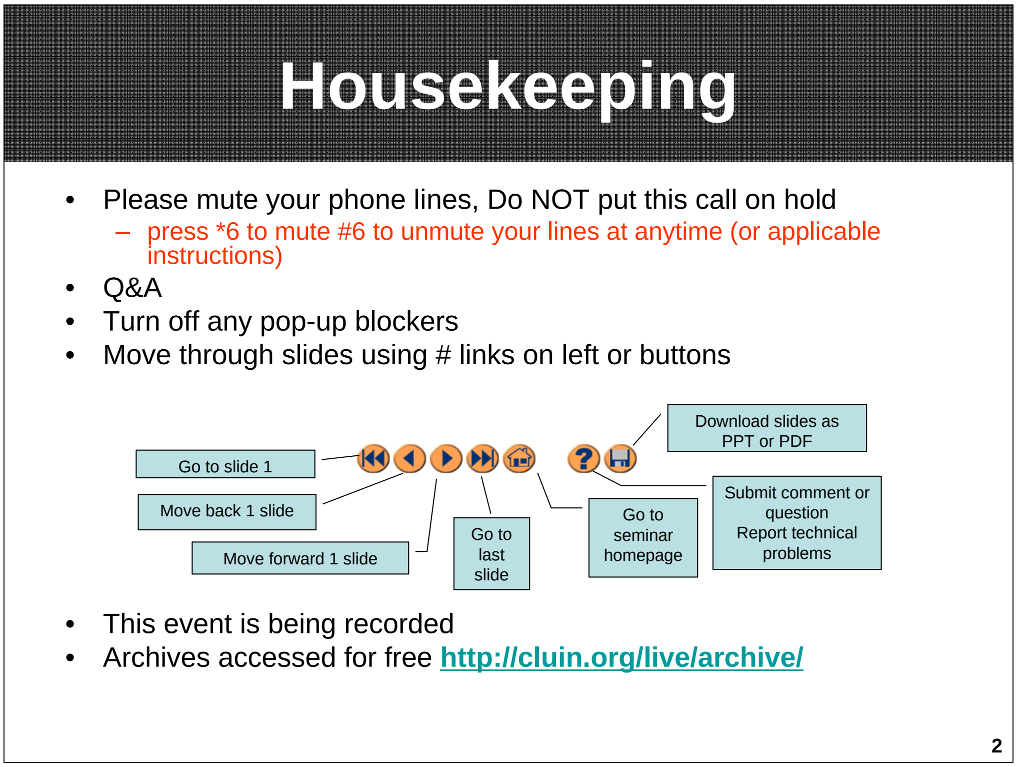
Kirk Cameron, MacStat Consulting, Ltd (kcmacstat@qwest.net)

Mike Gansecki, U.S. EPA Region 8 (gansecki.mike@epa.gov)

Moderator:

Jean Balent, U.S. EPA, Technology Innovation and Field Services Division (balent.jean@epa.gov)

Visit the Clean Up Information Network online at www.cluin.org



Housekeeping

- Please mute your phone lines, Do NOT put this call on hold
  - press *6 to mute #6 to unmute your lines at anytime (or applicable instructions)
- Q&A
- Turn off any pop-up blockers
- Move through slides using # links on left or buttons

Go to slide 1
Move back 1 slide
Move forward 1 slide
Go to last slide
Go to seminar homepage
Submit comment or question Report technical problems
Download slides as PPT or PDF

- This event is being recorded
- Archives accessed for free **http://cluin.org/live/archive/**

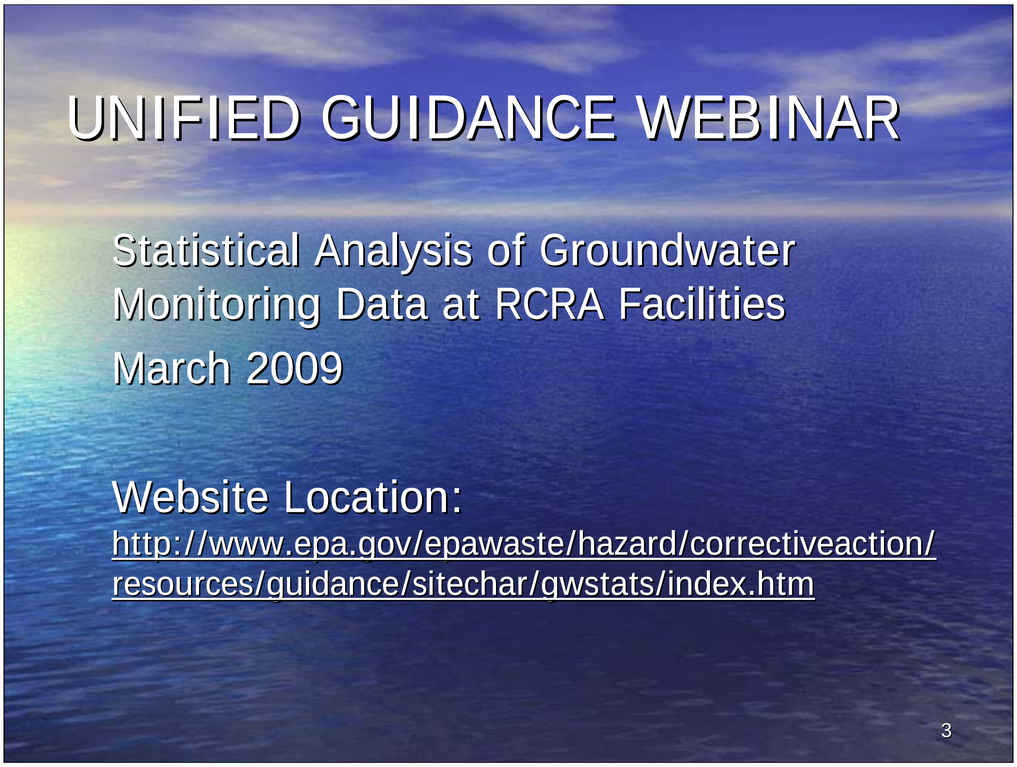
2

Although I’m sure that some of you have these rules memorized from previous CLU-IN events, let’s run through them quickly for our new participants.

Please mute your phone lines during the seminar to minimize disruption and background noise. If you do not have a mute button, press *6 to mute #6 to unmute your lines at anytime. Also, please do NOT put this call on hold as this may bring delightful, but unwanted background music over the lines and interupt the seminar.

You should note that throughout the seminar, we will ask for your feedback. You do not need to wait for Q&A breaks to ask questions or provide comments. To submit comments/questions and report technical problems, please use the ? Icon at the top of your screen. You can move forward/backward in the slides by using the single arrow buttons (left moves back 1 slide, right moves advances 1 slide). The double arrowed buttons will take you to 1st and last slides respectively. You may also advance to any slide using the numbered links that appear on the left side of your screen. The button with a house icon will take you back to main seminar page which displays our agenda, speaker information, links to the slides and additional resources. Lastly, the button with a computer disc can be used to download and save today’s presentation materials.

With that, please move to slide 3.



UNIFIED GUIDANCE WEBINAR

Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities

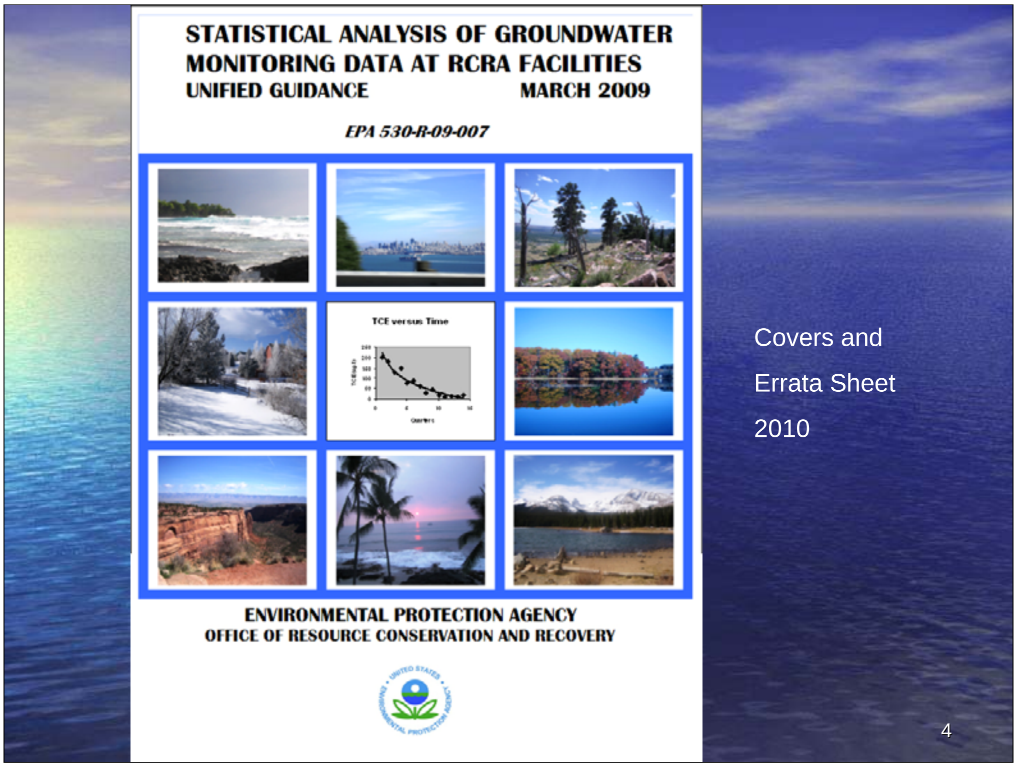
March 2009

Website Location:

http://www.epa.gov/epawaste/hazard/correctiveaction/resources/guidance/sitechar/gwstats/index.htm

3

Welcome to the EPA Webinar on the Unified Guidance. My name is Mike Gansecki of EPA Region 8. I served as the work assignment manager in completing this guidance. Formally, the guidance is titled the “Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities”, with a completion date of March 2009. It represents the combined efforts of statisticians, a workgroup, peer and other reviewers, the principal author Kirk Cameron of MacStat Corporation, covering a decade-long period. The document was approved and released in July 2009 on the EPA OSWER website shown on the slide.



STATISTICAL ANALYSIS OF GROUNDWATER MONITORING DATA AT RCRA FACILITIES

UNIFIED GUIDANCE MARCH 2009

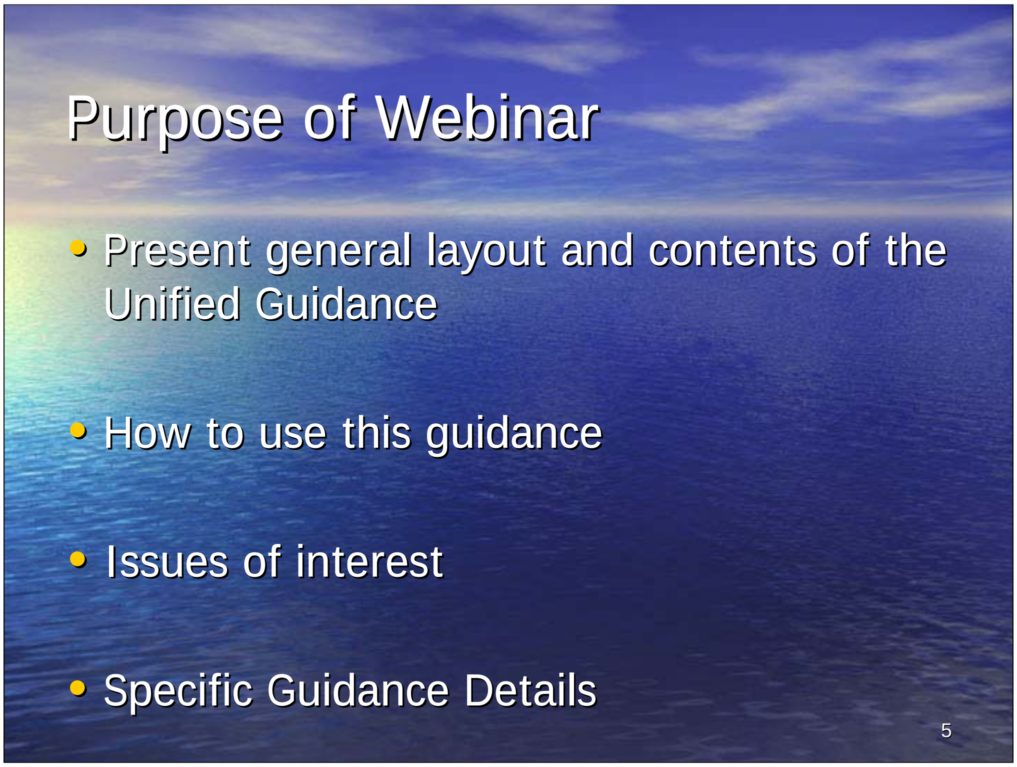
EPA 530-R-09-007

ENVIRONMENTAL PROTECTION AGENCY
OFFICE OF RESOURCE CONSERVATION AND RECOVERY

Covers and
Errata Sheet
2010

Since its release in 2009, the Unified Guidance has been slightly modified. This slide shows a cover which can be used for printing hard copies; a similar cover is available for the Appendices. In addition, an August 2010 Errata sheet has been developed, correcting certain numerical mistakes found in the guidance since its release. This Errata sheet has been placed on the EPA website along with corrected guidance files.

Statistics is not necessarily known for its visual attractiveness. This was the major reason for including these guidance cover photos of the America we are trying to protect. We also beg your indulgence in adding some color in the form of Wikipedia picture slides later to lighten and brighten this presentation.



Purpose of Webinar

- Present general layout and contents of the Unified Guidance
- How to use this guidance
- Issues of interest
- Specific Guidance Details

5

Covering the numerous statistical details of the 887 page Unified Guidace is impossible in a webinar. Today, we will discuss four basic topics. The first will be to present the general layout and contents found in the guidance. A second will provide some insights as to how the guidance can be used. Certain regulatory issues of interest will be briefly summarized. In the second portion of the webinar, Dr. Kirk Cameron will cover specific details of the guidance. There will be time for questions, once the presentation is finished. If we cannot directly provide answers, you will be notified where the questions and responses will be posted on a website at a later date.

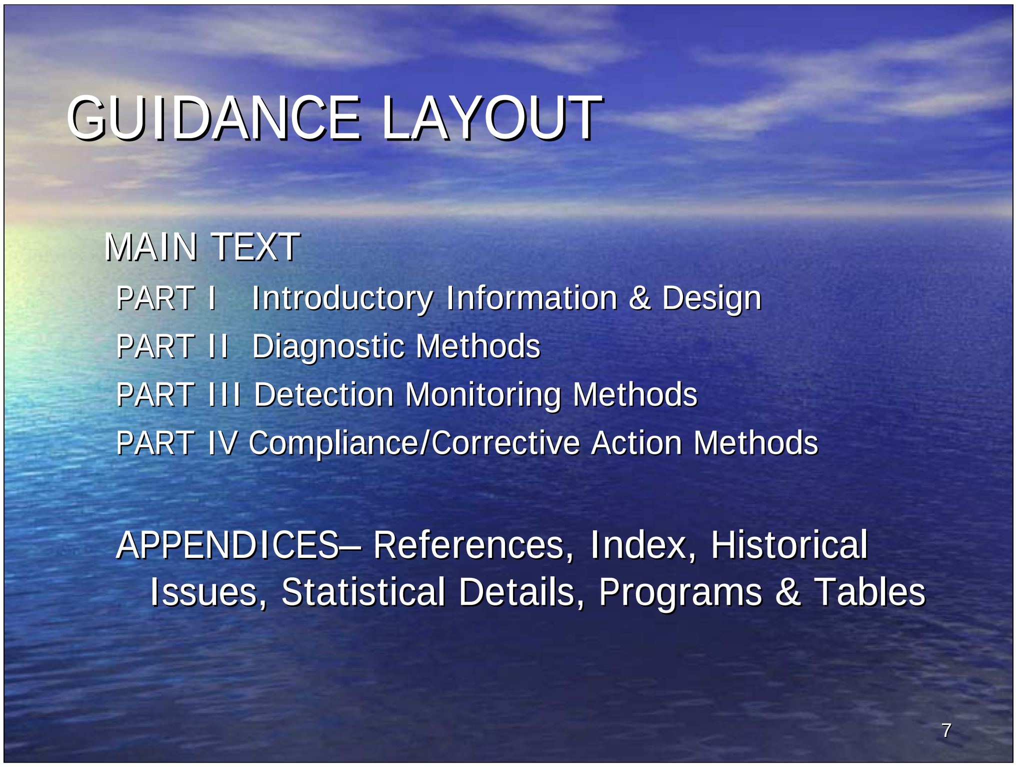
We reworked the guidance to flow more logically, with basic and general concepts at the outset, followed by more detailed methods and procedures in later Parts and Chapters of the guidance. While the guidance is written to cover both the Subtitle D solid waste and Subtitle C hazardous waste RCRA regulations, the approaches and methods can be applicable to other programs, such as Superfund. I hope you will find the information useful.

GENERAL LAYOUT

Longleat, England

6

Statistics can be daunting. Hopefully, the Unified Guidance layout will be easier to negotiate than the world’s longest hedge maze in Longleat, England!



GUIDANCE LAYOUT

MAIN TEXT

PART I Introductory Information & Design

PART II Diagnostic Methods

PART III Detection Monitoring Methods

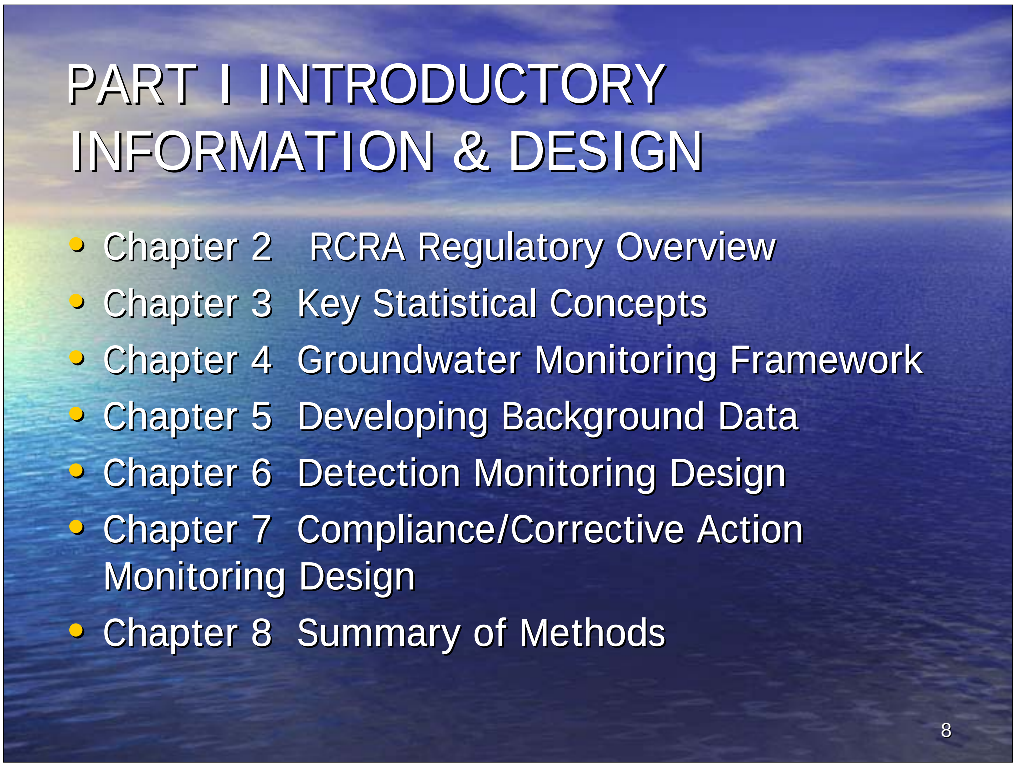
PART IV Compliance/Corrective Action Methods

APPENDICES– References, Index, Historical Issues, Statistical Details, Programs & Tables

The main text first covers introductory information in Part I, such as basic concepts and general design. We then include a set of methods and procedures for Diagnostic testing in Part II. Part III covers the RCRA regulatory formal statistical testing methods in a detection monitoring program, as well as other applicable and necessary approaches. Part IV similarly covers compliance and corrective action formal testing methods and strategies.

The appendices contain references, a glossary and index. Historical guidance discussions have also been moved to an appendix, as well as more detailed statistical calculations especially for power analyses. Three special programs have also been added, written in R-script (a public domain software). These can be used for difficult and tedious calculations for certain methods. Finally, there are extensive statistical tables to complement the various procedures and methods.

We chose procedures and methods which we believe can address most situations in the groundwater monitoring and testing context. However, the guidance is not intended to be exhaustive, and numerous references to other possible methods in the wider statistical literature are mentioned. We stress at the outset that this is only guidance, not regulation or policy.



PART I INTRODUCTORY INFORMATION & DESIGN

- Chapter 2 RCRA Regulatory Overview
- Chapter 3 Key Statistical Concepts
- Chapter 4 Groundwater Monitoring Framework
- Chapter 5 Developing Background Data
- Chapter 6 Detection Monitoring Design
- Chapter 7 Compliance/Corrective Action Monitoring Design
- Chapter 8 Summary of Methods

Part I introductory information first covers those portions of the RCRA groundwater monitoring regulations of concern from a statistical standpoint in Chapter 2. These include performance criteria, sampling size and frequency requirements, and certain historical issues raised during the long tenure of RCRA regulations dating back to 1980.

We have added Chapter 3 on basic statistical concepts, simple measures, definitions and assumptions. The latter include the need for statistical independence in many tests, stationarity, normality and other assumptions. Chapter 3 concepts and assumptions are used throughout the guidance.

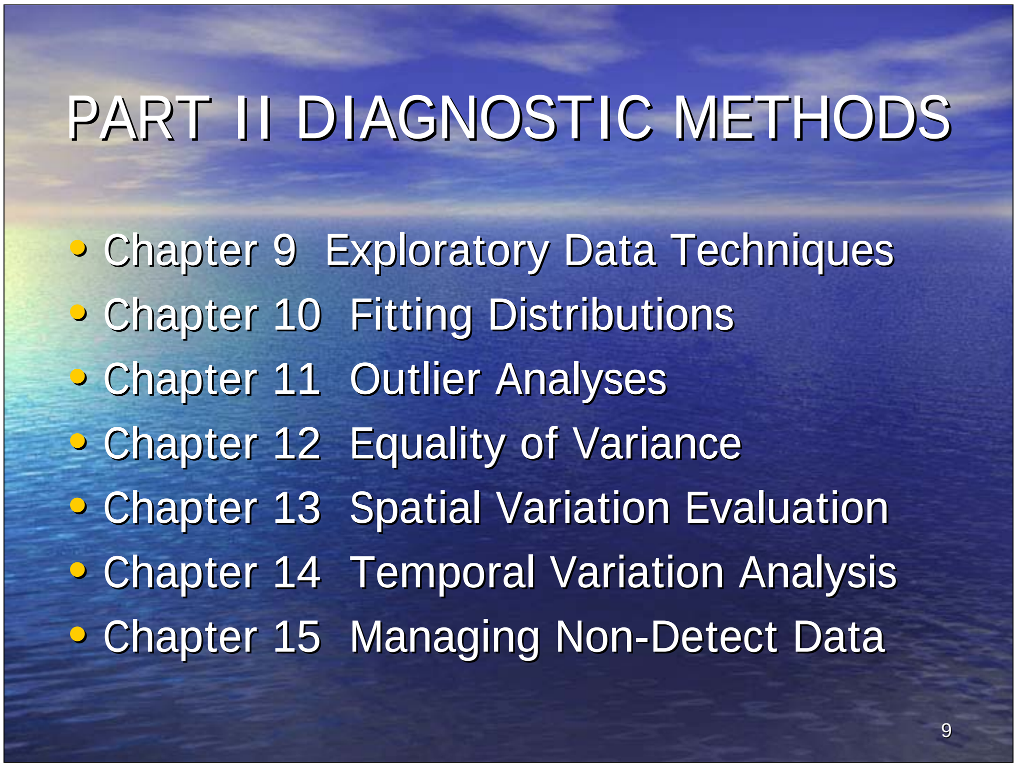
Chapter 4 lays out the basic structure of the groundwater monitoring and testing system. In this chapter, we also compiled a list of potential complicating factors when considering statistical significance for a given test.

Developing and updating background data has been of concern to both regulators and regulated parties. We look at approaches for developing adequate size background data sets (for example, in a permit), as well as periodically updating background information in Chapter 5.

Chapter 6 discusses detection monitoring design. These comparisons to background data are unique to each facility undergoing such development, and call for a systematic approach. We look at key topics in framing such a detection monitoring system. Because numerous constituents, compliance wells and annual tests are a common feature of detection monitoring, there is a strong need to control the overall false positive rate yet still maintain adequate power to detect releases. The guidance provides a structured and systematic approach which can be used at most, if not all, facilities developing or revising a detection monitoring program.

The following Chapter 7 considers compliance/corrective action monitoring design. In contrast to detection monitoring, comparisons to fixed health- or risk-based limits involve different testing approaches. These are not formally identified in regulation; hence the guidance offers a number of options. However, a regulatory agency will have a much greater say in identifying many of the appropriate aspects of such a system. These include the choice of statistical testing hypotheses and parameter(s), the appropriate limits, tolerable false positive and negative errors. When background limits are used for compliance monitoring, the guidance discusses reasonable ways to develop a system, including using the methods appropriate for detection monitoring (Section 7.5).

The final Chapter 8 of this Part summarizes each of the major methods found in the guidance, along with caveats and assumptions.



PART II DIAGNOSTIC METHODS

- Chapter 9 Exploratory Data Techniques
- Chapter 10 Fitting Distributions
- Chapter 11 Outlier Analyses
- Chapter 12 Equality of Variance
- Chapter 13 Spatial Variation Evaluation
- Chapter 14 Temporal Variation Analysis
- Chapter 15 Managing Non-Detect Data

Understanding the statistical properties of data is key to applying the potential tests found in the later Parts of the guidance. Both informal and formal diagnostic evaluation provides the means to identify those properties.

Initial Chapter 9 of Part II encourages the use of preliminary data analysis. Plotting data, visually identifying patterns with box plots, scatterplots, etc. are strongly encouraged here and throughout the guidance. While this Chapter includes typical methods, these are by no means the only options.

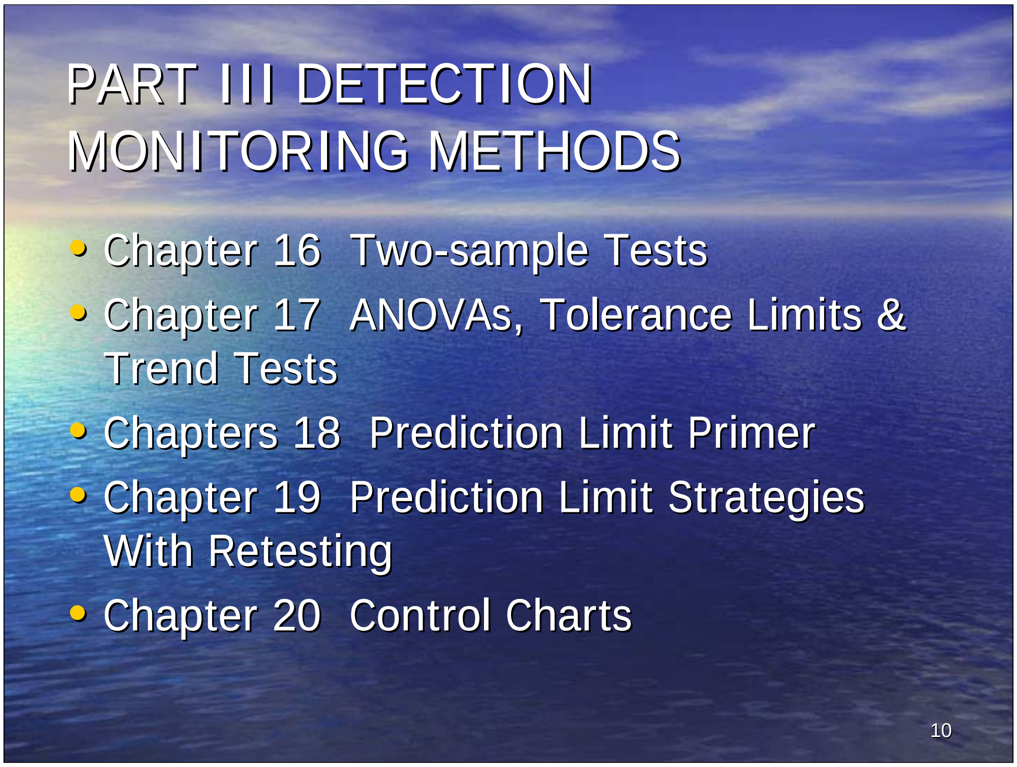
When data can be fitted to a parametric distribution, there are considerable advantages of efficiency using their mathematical properties. In Chapter 10, the guidance limits itself to consideration of the parametric family of normal distributions (normal, logarithmic, ladder-of-powers), but other options (particularly the gamma) are possible. Both informal and formal methods for evaluating distributional fit are provided.

The presence of outliers in data can negate otherwise useful applications. Chapter 11 includes two methods applicable to normalizable data. Other more robust methods are mentioned as possible applications, when normal tests are inappropriate.

Chapter 12 covers equality of variance testing. This assumption can be important for certain tests such as simple ANOVA. Chapter 13 evaluates potential spatial variation. Many commonly tested parameters (inorganic indicator constituents in particular), may exhibit well-to-well variation. Spatial variation can confound typical upgradient-to-downgradient well data testing, as well as in pooling multiple well data sets. ANOVA is used to diagnose well spatial variation, and a method is suggested for pooling certain data sets with unequal spatial means but with a common variance.

Chapter 14 looks at a number of temporal variation patterns and suggests certain remedies. These temporal patterns can include autocorrelation, trends, and seasonality in individual data sets, as well as co-variation of individual constituents at different wells, and multiple constituents in a single well.

Chapter 15 addresses management of non-detect data, including two new methods for evaluating and estimating non-detect data in multiple ND data sets.



PART III DETECTION MONITORING METHODS

- Chapter 16 Two-sample Tests
- Chapter 17 ANOVAs, Tolerance Limits & Trend Tests
- Chapters 18 Prediction Limit Primer
- Chapter 19 Prediction Limit Strategies With Retesting
- Chapter 20 Control Charts

10

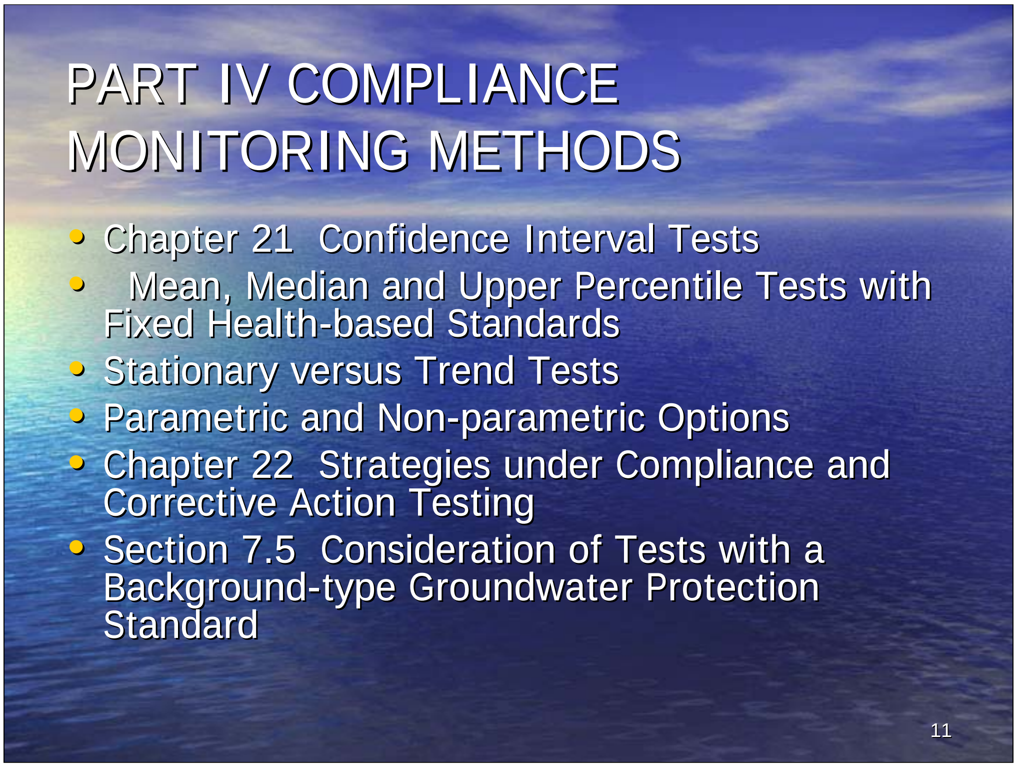
Part III covers the RCRA-specified formal detection monitoring testing options. We have provided non-parametric testing options in each chapter, when parametric assumptions cannot be met. Two-sample tests of downgradient well data versus background under Part 265 interim status regulations are covered in Chapter 16 of the guidance. This chapter also includes a Tarone-Ware test method for managing data with multiple detection limits. Except for the smallest of facilities with few monitoring constituents and wells, two-sample tests may not adequately maintain a low-enough false positive error rate.

Parametric and non-parametric ANOVA, tolerance intervals, prediction limits and control charts are the primary optional methods in the RCRA regulations. These classical tests require that data be stationary. If a trend is present, other methods for trend analysis are recommended for use.

Chapter 17 covers the two ANOVAs and tolerance interval options, as well as trend testing. ANOVAs used for formal detection monitoring are expected to have limited applicability, since many monitoring constituents exhibit spatial well variability. Trend tests (either parametric and non-parametric) are used when stationarity cannot be assumed.

Among the five RCRA tests, prediction limit theory (Chapters 18 and 19) is best developed to fully consider false positive and negative error rates in system design. For facilities with many annual detection monitoring statistical tests, the guidance offers a way to manage these error rates with reasonably minimum sample data sizes. Even a single future well observation can suffice as a test, when combined with conditional repeat sampling included with the testing design. Chapter 18 presents the basic parametric and non-parametric prediction limit tests. Chapter 19 provides nine parametric and six non-parametric prediction limit tests involving repeat samples to address false positive error rate and power design concerns. A stand-alone Optimal Rank Calculator also accompanies this guidance to allow the optimal choice of the maximal value for certain non-parametric prediction limit tests of future observations.

Similar results and performance can also be attained with control charts, as described in Chapter 20. The Shewhart-Cusum combined control chart test is recommended as having superior performance. Both prediction limits and control charts do require stationary data. Control chart performance also depends on an assumption of normality. It too can make use of designed repeat sampling to enhance performance goals.



PART IV COMPLIANCE MONITORING METHODS

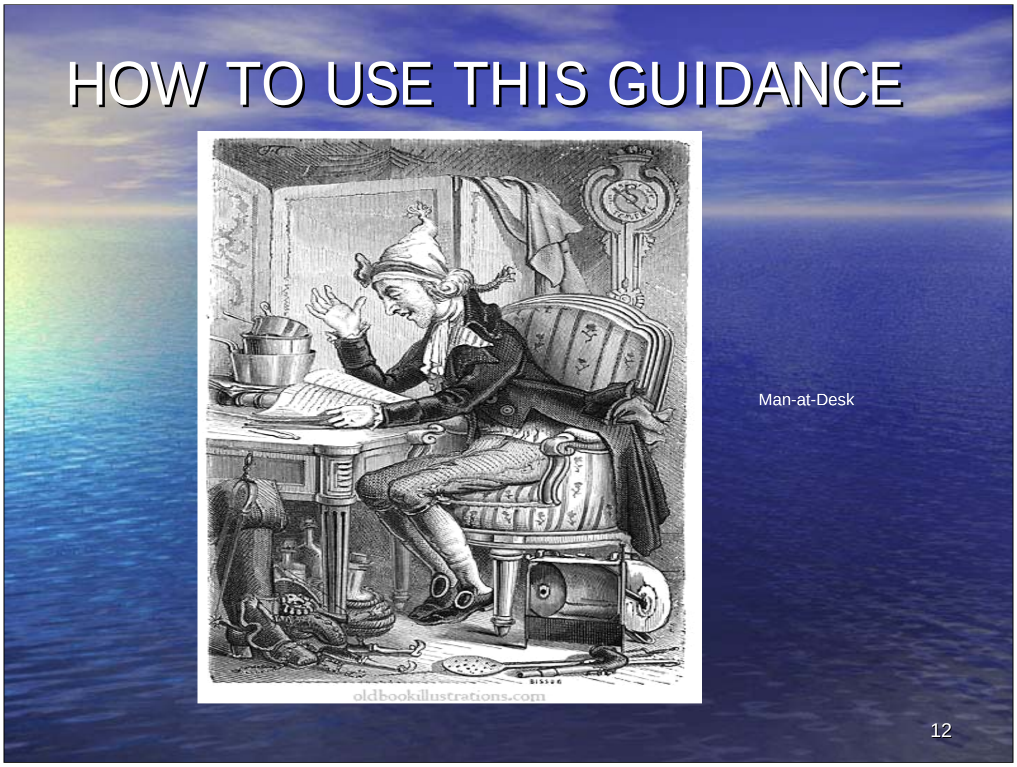
- Chapter 21 Confidence Interval Tests
- Mean, Median and Upper Percentile Tests with Fixed Health-based Standards
- Stationary versus Trend Tests
- Parametric and Non-parametric Options
- Chapter 22 Strategies under Compliance and Corrective Action Testing
- Section 7.5 Consideration of Tests with a Background-type Groundwater Protection Standard

Part IV covers compliance and corrective action formal tests against a fixed health- or risk-based limit. Chapter 21 identifies the principal types of tests, while Chapter 22 provides strategies for statistical design. A reading of Chapter 7 is also important in understanding the general design principles and context.

Comparisons against a fixed limit or Groundwater Protection Standard (GWPS) make use of an upper or lower confidence interval comparison (depending on the null hypothesis for compliance and corrective action testing). As noted in Chapter 7, a regulatory decision must be first made regarding the appropriate statistical parameter for comparison as well as appropriate false positive and negative error criteria. For parametric tests, the parameter may be the normal mean, logarithmic (arithmetic or geometric) mean or some upper percentile. If sample data indicate a trend, a linear regression approach is suggested. Similarly, non-parametric versions of these tests are provided (median, upper percentile or median trend evaluation tests).

In compliance monitoring statistical design, the guidance concludes that a different false positive and power control approach should be used, than for detection monitoring background comparisons. For compliance testing, required power is first identified, and the false positive error rate for a single well-constituent test can then be adjusted. In contrast for corrective action, a fixed single test error rate is selected, and sample sizes can be identified which can meet pre-set power criteria. The guidance also suggests pooling of historical data to enhance sample sizes.

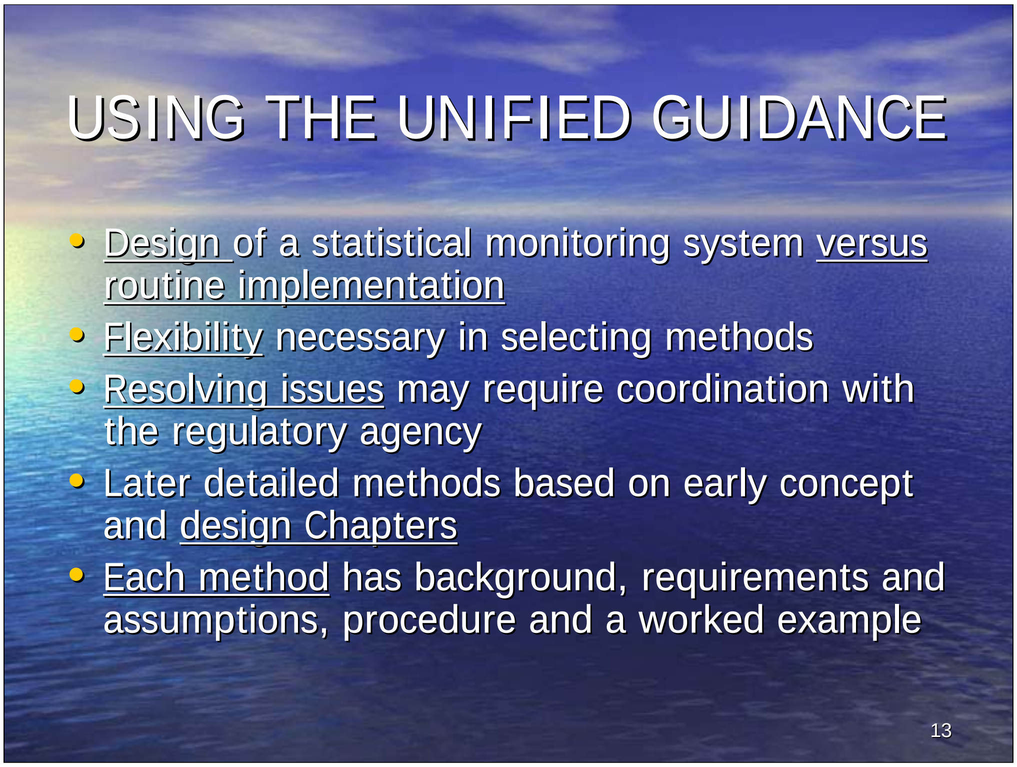
If a background-type GWPS is used, two different approaches are suggested (in Section 7.5). If a single-sample compliance test is used, the strategies suggested in Chapters 21 and 22 can be used. If two- or multiple samples will be compared, the tests and strategies suggested for detection monitoring (Chapters 6, 16-20) can be used. A mini-max strategy may also be considered, given data size limitations (Section 7.5).



HOW TO USE THIS GUIDANCE

Man-at-Desk

While the document may seem overwhelming to use at times, the Eureka Moment comes when you learn and apply the right tools. This gentleman has already acquired many of the necessary tools.



USING THE UNIFIED GUIDANCE

- Design of a statistical monitoring system versus routine implementation
- Flexibility necessary in selecting methods
- Resolving issues may require coordination with the regulatory agency
- Later detailed methods based on early concept and design Chapters
- Each method has background, requirements and assumptions, procedure and a worked example

13

Much of the Unified Guidance is oriented towards development of the statistical aspects of one or more monitoring systems (detection, compliance, or corrective action). This would occur, for example, in generating a permit, order, or periodically modifying an existing system. While the guidance offers numerous alternative methods, a decision must select only one method for final routine implementation. The flexibility offered in the guidance is necessary given very different data patterns, the type of monitoring system, regulatory and regulated facility needs, etc. In particular, detection monitoring decisions will be mostly site-specific, depending on the well monitoring system, number and types of monitoring constituents, and their prior history and behavior. In contrast, compliance and corrective action monitoring will generally involve hazardous constituents and pre-set GWPS. In the latter cases, decisions about the eventual system need to be closely coordinated with regulatory agency policies and determinations.

Users should first become familiar with the basic and general concepts of the first Part of the guidance, especially regarding the relevant statistical design concepts and issues. Selection of appropriate methods can then follow, based to a great extent on constituent data patterns and regulatory agency decisions or policies.

The guidance methods (summarized in Chapter 8) for diagnostic as well as formal testing are laid out in a consistent fashion. For each method, there is a brief discussion of the principles, background and purpose, as well as relevant assumptions and constraints. This is then followed by a step-wise procedure. Finally, one or more worked examples are provided for each method. In the slides which follow, we will look at the information provided for one diagnostic method, the Rank vonNeumann Ratio test for autocorrelation (temporal variation) in a data set.

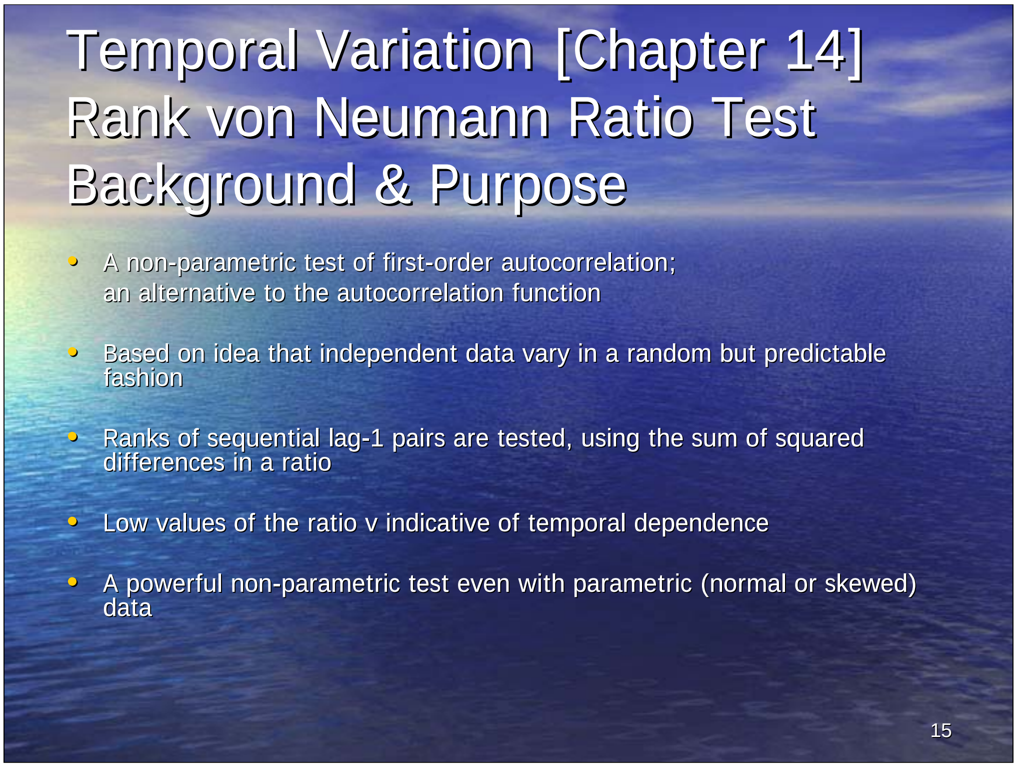
We will then consider the arsenic data in this method example using other relevant diagnostic information from Part II. Once these patterns have been established, the data set can be considered for use with one or more of the formal monitoring tests in Part III or IV.

The Neumanns

Alfred E. Neuman, Cover of MAD #30

John von Neumann, taken in the 1940's

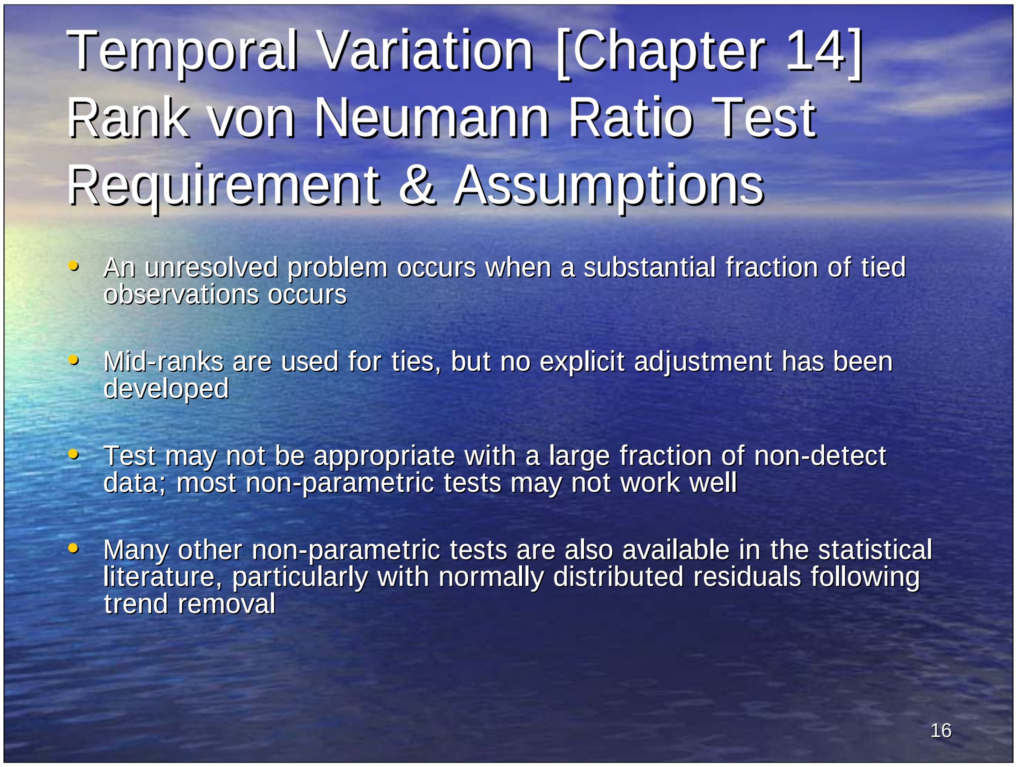
Just to be clear--- This is not the “What-me-worry” test of MAD magazine’s Alfred E. Neuman. Rather it is one developed by John von Neumann, a brilliant Hungarian-American mathematician. He’s not only famous for his work in statistics, but in game theory, theory of automata, and quantum physics as well. He was also a major participant on the Manhattan Project during World War II.



Temporal Variation [Chapter 14] Rank von Neumann Ratio Test Background & Purpose

- A non-parametric test of first-order autocorrelation; an alternative to the autocorrelation function
- Based on idea that independent data vary in a random but predictable fashion
- Ranks of sequential lag-1 pairs are tested, using the sum of squared differences in a ratio
- Low values of the ratio v indicative of temporal dependence
- A powerful non-parametric test even with parametric (normal or skewed) data

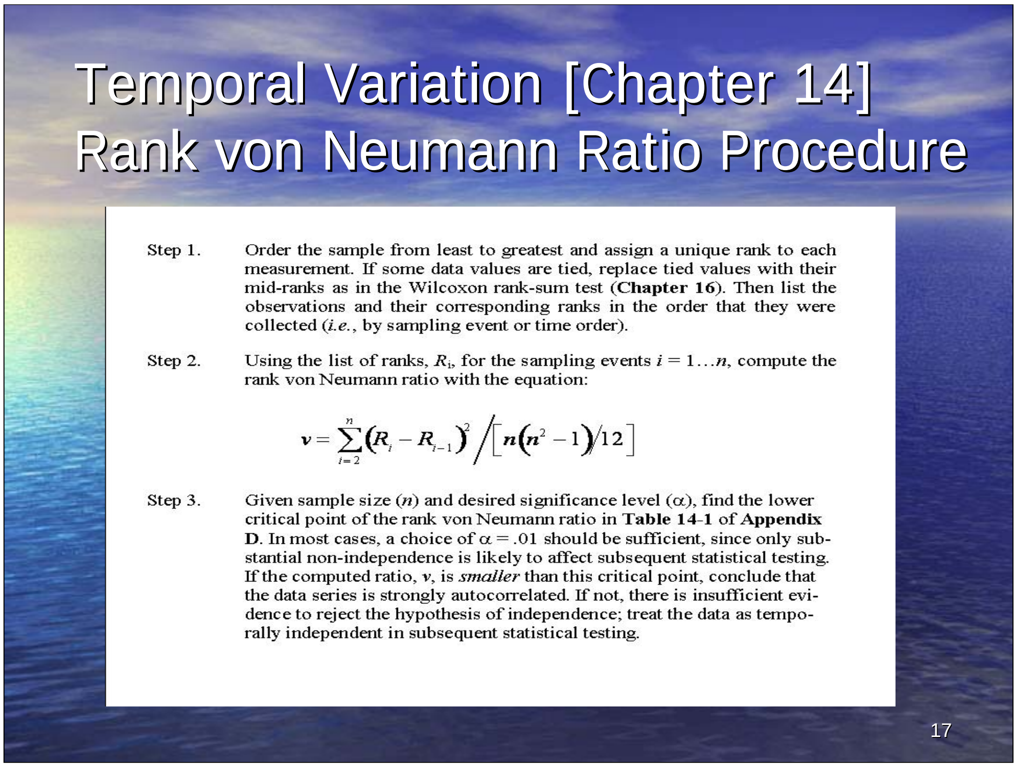
BACKGROUND AND PURPOSE: The Rank vonNeumann Ratio method in Section 14.2.4 is a test of first-order autocorrelation (i.e., between successive data in time), an alternative to the autocorrelation function. The basic principle is that independent data will vary in a random but predictable fashion, while autocorrelated dependent data will more systematically vary. The method uses rankings of sequential, single time-lag data, to provide a sum of squared differences in a ratio measurement. Low values are indicative of many forms of temporal dependence. It is a powerful test even as applied to parametric normal or skewed data sets.



Temporal Variation [Chapter 14] Rank von Neumann Ratio Test Requirement & Assumptions

- An unresolved problem occurs when a substantial fraction of tied observations occurs
- Mid-ranks are used for ties, but no explicit adjustment has been developed
- Test may not be appropriate with a large fraction of non-detect data; most non-parametric tests may not work well
- Many other non-parametric tests are also available in the statistical literature, particularly with normally distributed residuals following trend removal

REQUIREMENTS AND ASSUMPTIONS: One constraint in applying the Rank vonNeumann Ratio test is that a substantial fraction of tied values or observations result in an uncertain outcome. For a limited number of ties, mid-rankings are used, but there has not been any explicit adjustment developed to account for these ties. A similar situation also occurs when there is a large fraction of non-detect data. Most non-parametric tests will not perform suitably under these conditions. The guidance notes that there are many other non-parametric tests of autocorrelation in the wider statistical literature, especially where normally distributed residuals occur following trend removal.



Temporal Variation [Chapter 14] Rank von Neumann Ratio Procedure

Step 1. Order the sample from least to greatest and assign a unique rank to each measurement. If some data values are tied, replace tied values with their mid-ranks as in the Wilcoxon rank-sum test (**Chapter 16**). Then list the observations and their corresponding ranks in the order that they were collected (*i.e.*, by sampling event or time order).

Step 2. Using the list of ranks, R_i, for the sampling events $i = 1 \ldots n$, compute the rank von Neumann ratio with the equation:

$$\boldsymbol{v} = \sum_{i=2}^{n} \left(R_i - R_{i-1}\right)^2 \Big/ \left[\boldsymbol{n}\left(\boldsymbol{n}^2 - 1\right)/12\right]$$

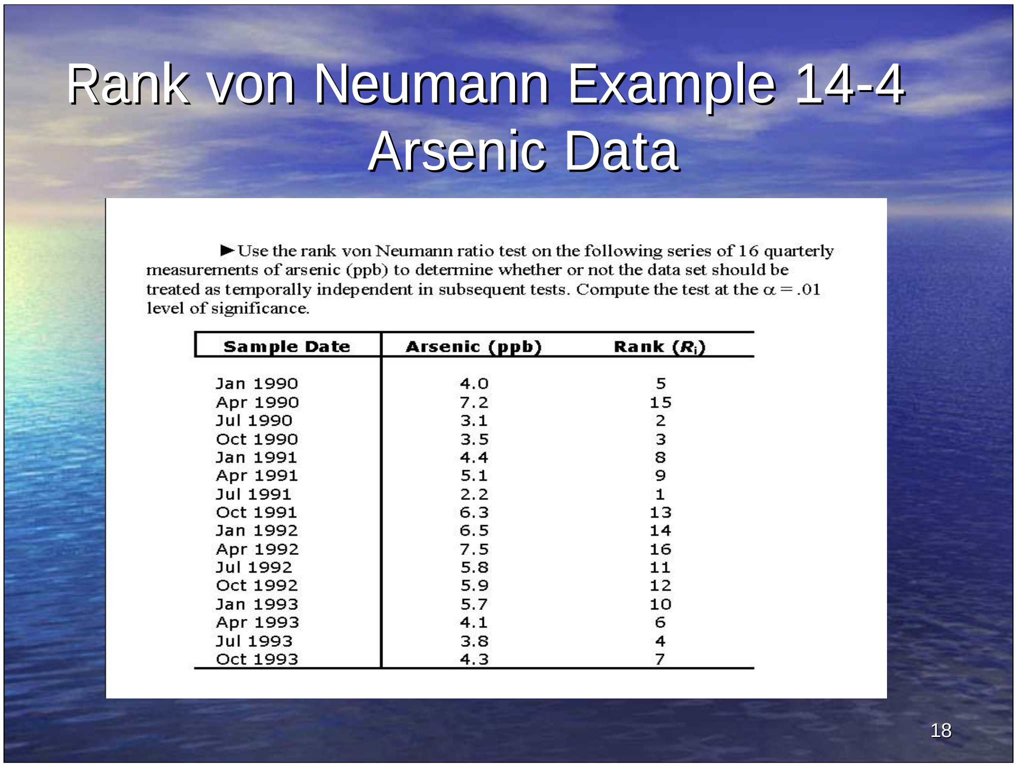
Step 3. Given sample size (n) and desired significance level (α), find the lower critical point of the rank von Neumann ratio in **Table 14-1** of **Appendix D**. In most cases, a choice of $\alpha = .01$ should be sufficient, since only substantial non-independence is likely to affect subsequent statistical testing. If the computed ratio, v, is *smaller* than this critical point, conclude that the data series is strongly autocorrelated. If not, there is insufficient evidence to reject the hypothesis of independence; treat the data as temporally independent in subsequent statistical testing.

17

PROCEDURE: The following three steps identify how the ranking ratio is calculated. In the first step, order the data in increasing size, and identify the ranks. In the presence of ties, replace tied values with their mid-rankings. Then list the observations and their rankings in original time event order.

In Step 2, calculate (Greek letter) nu as the sum of successive squared rank differences divided by a constant denominator [n x (n squared -1)/12]. The denominator is specific to the sample size n for independent data.

Compare the nu value for a desired significance level to the appropriate critical point from Table 14-1 in Appendix D. Typically, a .01 significance level is used. If nu is smaller than this critical point, one can conclude that there is strong evidence of first-order autocorrelation. Otherwise, the data can be treated as independent.



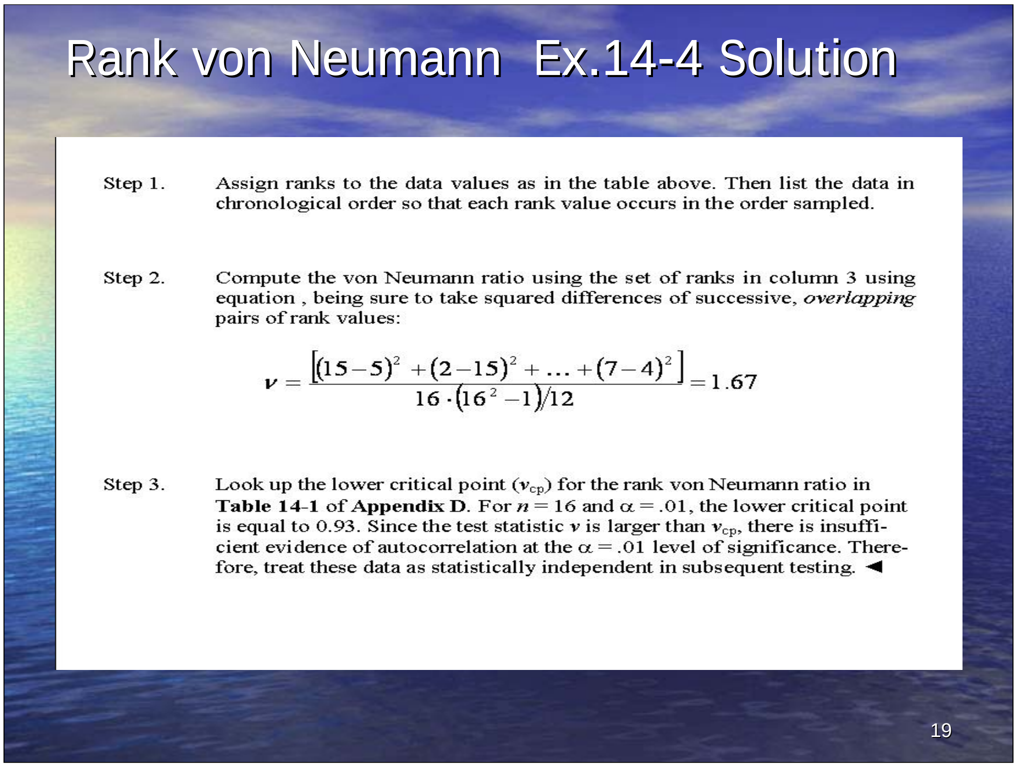
Rank von Neumann Example 14-4 Arsenic Data

►Use the rank von Neumann ratio test on the following series of 16 quarterly measurements of arsenic (ppb) to determine whether or not the data set should be treated as temporally independent in subsequent tests. Compute the test at the $\alpha = .01$ level of significance.

Sample Date	Arsenic (ppb)	Rank (R_i)
Jan 1990	4.0	5
Apr 1990	7.2	15
Jul 1990	3.1	2
Oct 1990	3.5	3
Jan 1991	4.4	8
Apr 1991	5.1	9
Jul 1991	2.2	1
Oct 1991	6.3	13
Jan 1992	6.5	14
Apr 1992	7.5	16
Jul 1992	5.8	11
Oct 1992	5.9	12
Jan 1993	5.7	10
Apr 1993	4.1	6
Jul 1993	3.8	4
Oct 1993	4.3	7

18

EXAMPLE: In Example 14-4 illustrating this procedure, the arsenic data shown in the table are provided as reported over time. The table also includes the rank values associated with each arsenic event data point. Note that there are no ties in this data set. A .01 significance level for the test is chosen.



Rank von Neumann Ex.14-4 Solution

Step 1. Assign ranks to the data values as in the table above. Then list the data in chronological order so that each rank value occurs in the order sampled.

Step 2. Compute the von Neumann ratio using the set of ranks in column 3 using equation , being sure to take squared differences of successive, *overlapping* pairs of rank values:

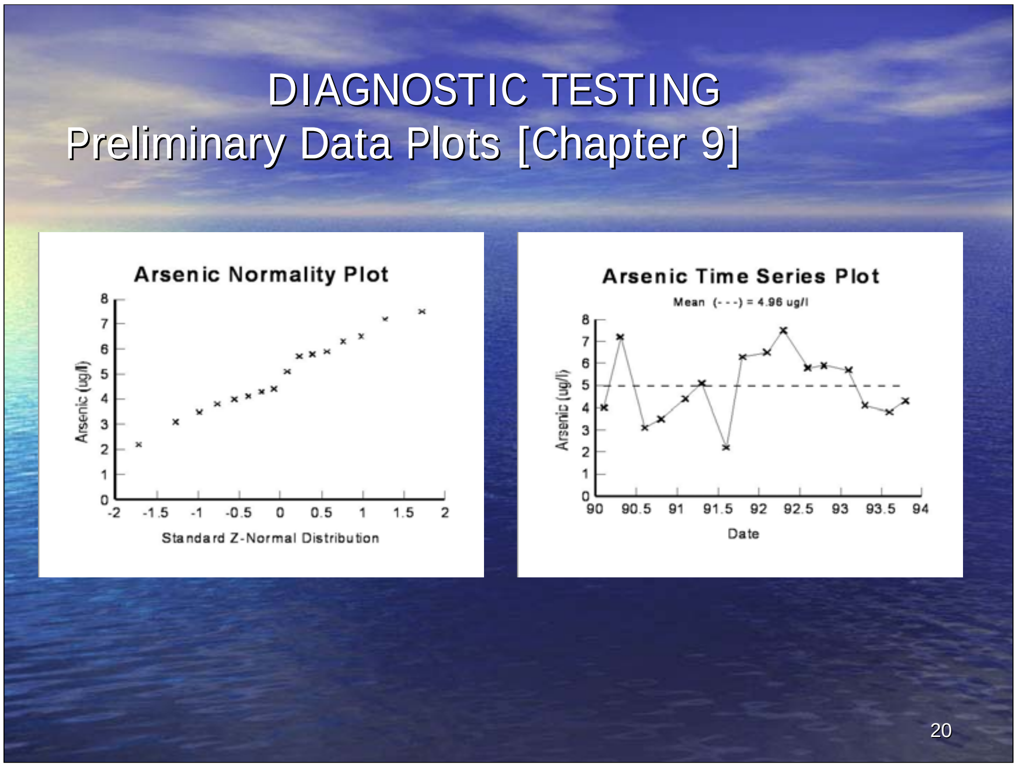
$$\nu = \frac{\left[(15-5)^2 + (2-15)^2 + \ldots + (7-4)^2\right]}{16\cdot\left(16^2-1\right)/12} = 1.67$$

Step 3. Look up the lower critical point (ν_{cp}) for the rank von Neumann ratio in **Table 14-1** of **Appendix D**. For $n = 16$ and $\alpha = .01$, the lower critical point is equal to 0.93. Since the test statistic ν is larger than ν_{cp}, there is insufficient evidence of autocorrelation at the $\alpha = .01$ level of significance. Therefore, treat these data as statistically independent in subsequent testing. ◄

Step 1 of the procedure has already been followed, as shown in the table. The computed nu value is obtained by taking the sum of the squared differences in the second to first rankings, the third to second, etc. and dividing by the constant denominator. The nu value is calculated as 1.67.

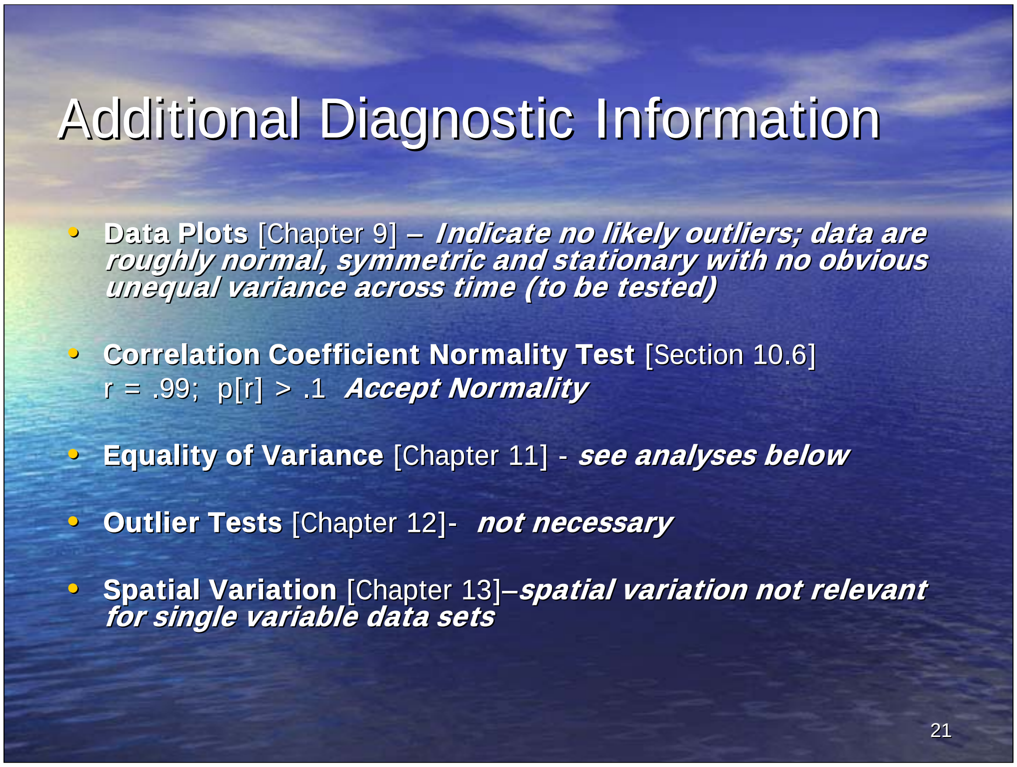
In Step 3, the critical value obtained from Table 14-1 for a .01 significance level and 16 data points is .93. Since the test nu value is greater, conclude that there is insufficient evidence of autocorrelation, and the data can be treated as statistically independent.

While this test provides good diagnostic information regarding autocorrelation/temporal independence, suppose the example arsenic data were under consideration as background data for a detection monitoring program. The following slides will summarize a wider set of diagnostic information and tests that can help determine which detection monitoring tests might be appropriate.



DIAGNOSTIC TESTING
Preliminary Data Plots [Chapter 9]

Other preliminary data evaluations from Chapter 9 can be used. Two very common plots are the ordered data versus a standard normal distribution, and a temporal plot of arsenic data against the sampling time events. In the second graph, the sample mean of 4.96 ug/l arsenic is also shown as a reference point for the time series plot. Data seem roughly symmetric around the sample mean. The normal probability plot suggests that the data may well be met by a normal distribution assumption. The normality plot also does not suggest extreme or even one or more likely outliers in this data set. The time series plot does not exhibit any obvious trend in the data, and the overall variance seems reasonably constant across time. These visual observations can be further evaluated formally using various Unified Guidance tests.



Additional Diagnostic Information

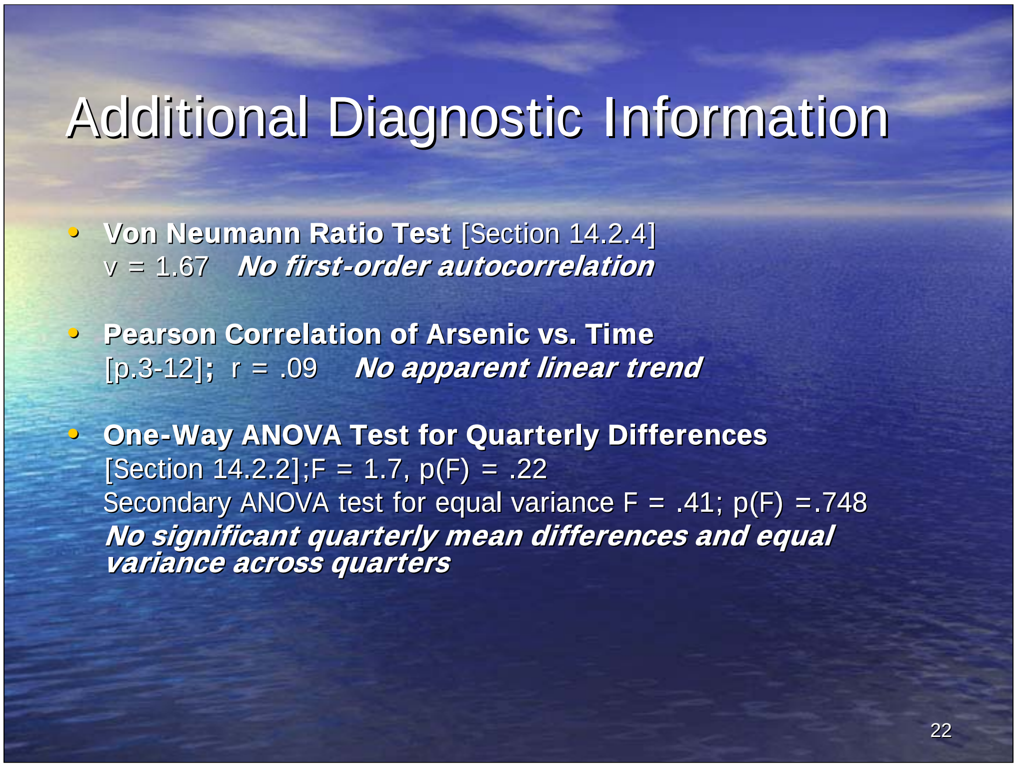
- **Data Plots** [Chapter 9] – ***Indicate no likely outliers; data are roughly normal, symmetric and stationary with no obvious unequal variance across time (to be tested)***
- **Correlation Coefficient Normality Test** [Section 10.6]
  r = .99; p[r] > .1 ***Accept Normality***
- **Equality of Variance** [Chapter 11] - ***see analyses below***
- **Outlier Tests** [Chapter 12]- ***not necessary***
- **Spatial Variation** [Chapter 13]–***spatial variation not relevant for single variable data sets***

21

In the present slide and ones which follow, we will consider each Part II diagnostic chapter of the guidance in succession. Although details of the procedures and calculations for these additional diagnostic tests are not shown, the conclusions are. Using the correlation coefficient test of Section 10.6, the r = .99 value allows us to accept a normal distribution assumption and fit for these data.

Equality of variance (Chapter 11) is considered for certain patterns under Chapter 14 Temporal Variation. We have already concluded that there is no need for formal outlier tests from Chapter 12, although one could run Dixon’s outlier test for the minimum or maximum values (assuming normal data).

Spatial variation (Chapter 13) is not a consideration for a single well data set, although it would be if multiple well data were considered for pooling.

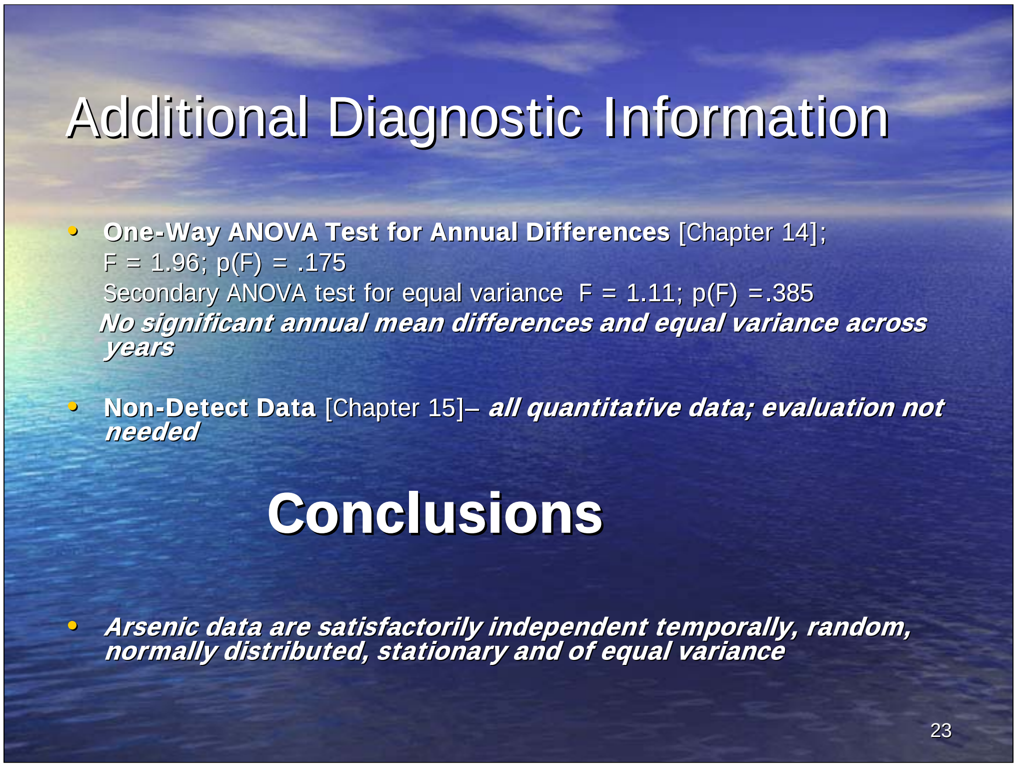


Additional Diagnostic Information

- **Von Neumann Ratio Test** [Section 14.2.4]
  v = 1.67 ***No first-order autocorrelation***
- **Pearson Correlation of Arsenic vs. Time**
  [p.3-12]; r = .09 ***No apparent linear trend***
- **One-Way ANOVA Test for Quarterly Differences**
  [Section 14.2.2];F = 1.7, p(F) = .22
  Secondary ANOVA test for equal variance F = .41; p(F) =.748
  No significant quarterly mean differences and equal variance across quarters

In addition to the results of the Rank von Neumann Ratio test, a Pearson linear correlation of arsenic values with time resulted in a correlation coefficient r = .09, clearly insignificant. This is further evidence of a lack of linear trend. If the data were not normally distributed or a transformation were required, the Mann-Kendall test of a significant median trend could be used. Because this test is also provided as a formal detection monitoring test involving a trend, it is found in Chapter 17, rather than in Part II diagnostic tests.

In order to check for either seasonal temporal variation or annual changes, a one-way ANOVA can be used. Secondary ANOVA tests of absolute residuals can be used to check equality of variance. For quarterly differences in the arsenic data set, there is no significant mean difference nor indications of unequal variance.



Additional Diagnostic Information

- **One-Way ANOVA Test for Annual Differences** [Chapter 14];
  F = 1.96; p(F) = .175
  Secondary ANOVA test for equal variance F = 1.11; p(F) =.385
  No significant annual mean differences and equal variance across years
- **Non-Detect Data** [Chapter 15]– ***all quantitative data; evaluation not needed***

Conclusions

- ***Arsenic data are satisfactorily independent temporally, random, normally distributed, stationary and of equal variance***

23

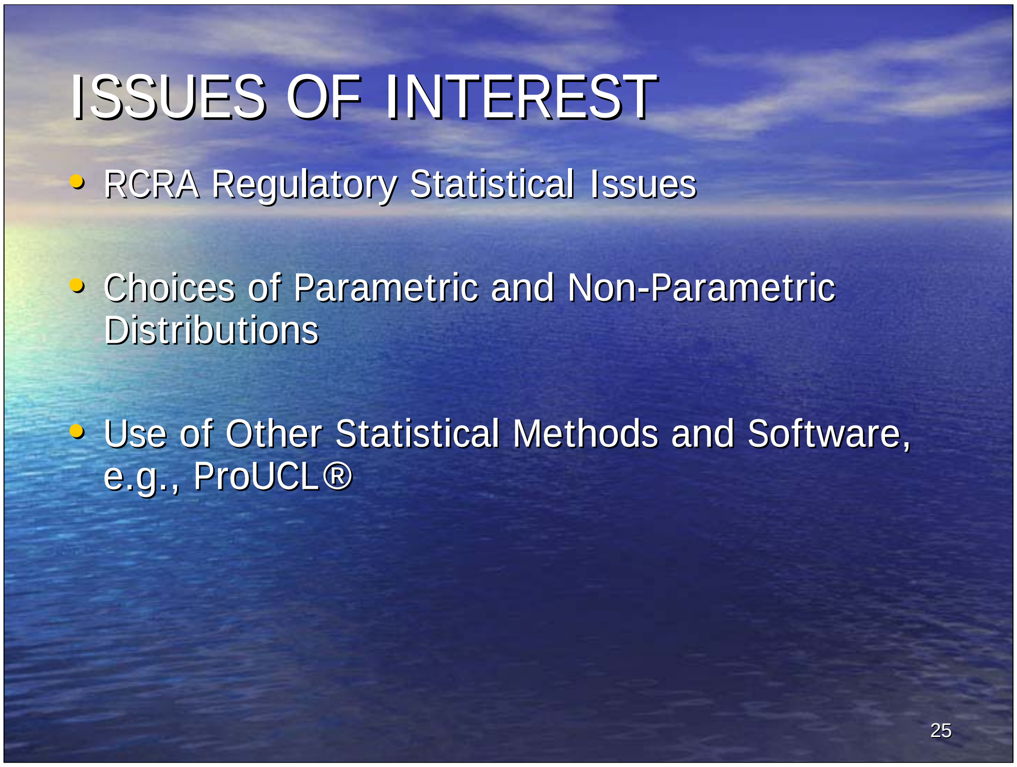
In similar fashion, the same ANOVA tests do not indicate any annual mean differences or unequal variance across years. Since the data are all quantitative, Chapter 15 non-detect management issues don't arise.

The overall conclusions from these diagnostic tests are that this arsenic data set can be assumed to be independent and random, normally distributed, stationary and of equal variance across time. Most of the parametric normal detection monitoring tests of Part III could be used with these arsenic data as background. Although this arsenic data set is relatively straightforward to diagnose, hopefully, these slides illustrate that there is a great deal of flexibility in how the Unified Guidance diagnostic tests can and should be used.

ISSUES

The Thinker, Musee Rodin in Paris

Pondering and interpreting thorny issues can leave one almost feeling as stumped as Rodin’s Thinker, but hopefully the guidance can help resolve some old, long-standing problems.

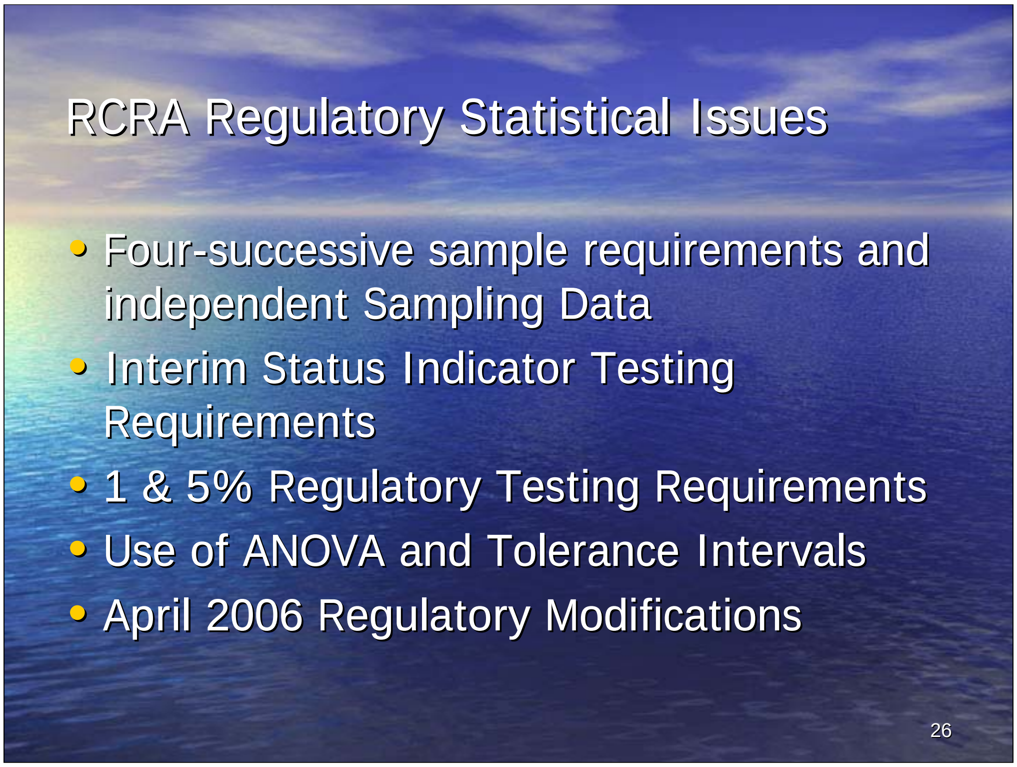


ISSUES OF INTEREST

- RCRA Regulatory Statistical Issues
- Choices of Parametric and Non-Parametric Distributions
- Use of Other Statistical Methods and Software, e.g., ProUCL®

A number of issues dealing with RCRA statistical applications have arisen over the years; some have been raised during the review phases of this guidance development. The Unified Guidance provides suggestions for addressing these issues while meeting the intent and spirit of the rules. Others are special features of this guidance, which need further explanation.

Topics Include certain RCRA regulatory statistical issues, using parametric and non-parametric distribution alternatives, and finally consideration of other methods in the statistical literature and available software. Each will be briefly discussed in turn.



RCRA Regulatory Statistical Issues

- Four-successive sample requirements and independent Sampling Data
- Interim Status Indicator Testing Requirements
- 1 & 5% Regulatory Testing Requirements
- Use of ANOVA and Tolerance Intervals
- April 2006 Regulatory Modifications

#1 The Unified Guidance takes the position that RCRA regulatory changes adopted in 1988 require independent sampling. We also indicate that generating and analyzing 4 aliquots of a single sample as required in the 1982 rules (since modified) will almost invariably violate the independence criterion and adversely affect test outcomes. We recommend physically separate samples collected with some minimum time between to ensure independence, where allowed. The guidance recognizes that existing state RCRA regulations may still require four successive samples for testing.

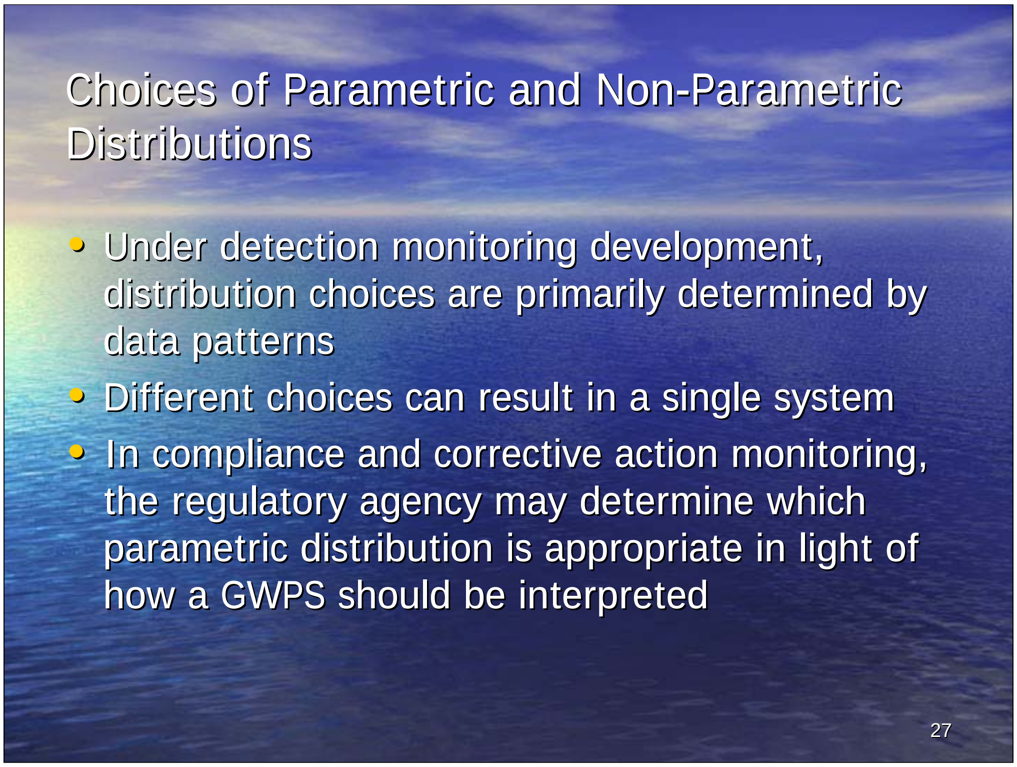
#2 For interim status monitoring, the guidance recommends moving from the four indicator tests in favor of a groundwater quality assessment monitoring plan to be developed under 40 CFR 265.90 or 93. Under the latter, a more flexible and realistic plan can be developed (including monitoring for hazardous constituents appropriate to Part 264 permits beyond those minimally required in Part 265).

#3 The guidance interprets the false positive requirements under 40 CFR 264.97(i) and 258.53(h) performance standards to apply to multiple sample tests such as t-tests, ANOVA, confidence interval compliance tests, but not for tolerance intervals, prediction limits and control charts.

#4 While the guidance provides procedures for ANOVA and tolerance interval detection monitoring tests, we also conclude that formal ANOVA-type tests will generally not be suitable when there is well spatial variation present. Tolerance intervals can still be used, but the guidance suggests that prediction limit theory is better established and would be a preferable alternative

#5 Chapter 2 contains a summary discussion of the principal 2006 changes to the Part 264 regulations. In addition to limiting the mandatory four-successive sampling requirement in favor of site-specific decisions, these 2006 rule changes also allow for more specific targeting of Appendix VIII/IX constituents required under certain circumstances as well as the compliance monitoring wells to be analyzed. Authorized State programs would need to adopt these changes to be applicable.

These interpretations may be different than applied under current State RCRA programs. If there is any doubt, the State regulatory agency should be contacted.



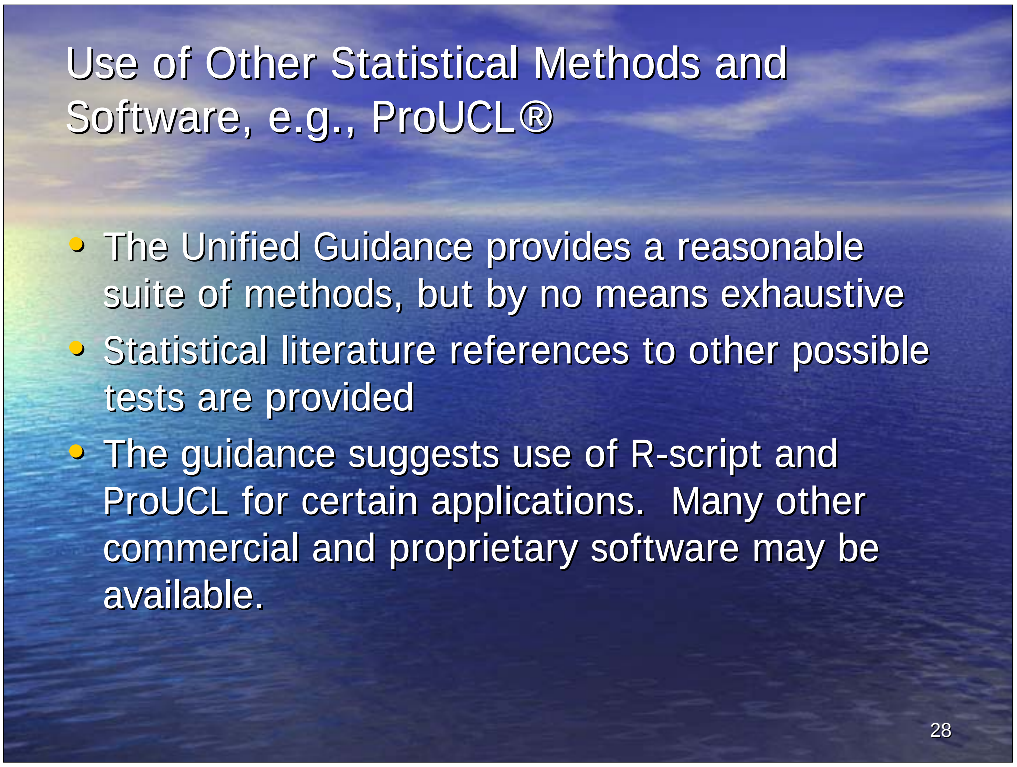
Choices of Parametric and Non-Parametric Distributions

- Under detection monitoring development, distribution choices are primarily determined by data patterns
- Different choices can result in a single system
- In compliance and corrective action monitoring, the regulatory agency may determine which parametric distribution is appropriate in light of how a GWPS should be interpreted

27

The Unified Guidance provides parametric (generally the normal distribution family) and non-parametric alternatives for most methods. In assessing background data sets in detection monitoring program development, the choices are largely determined by diagnostic test outcomes from Part II. It is likely that for a suite of well constituents, a number of different choices will be made (normal, lognormal, non-parametric). While the guidance does not provide other parametric distribution choices, these may also be obtained from outside sources and may prove superior.

In compliance/corrective action testing versus a fixed health- or risk-based limit, regulatory agency interpretations of the GWPS may limit certain choices. If an arithmetic mean is determined to best represent an MCL, for example, one or another normal tests will be needed. The guidance offers suggestions and caveats in this regard.



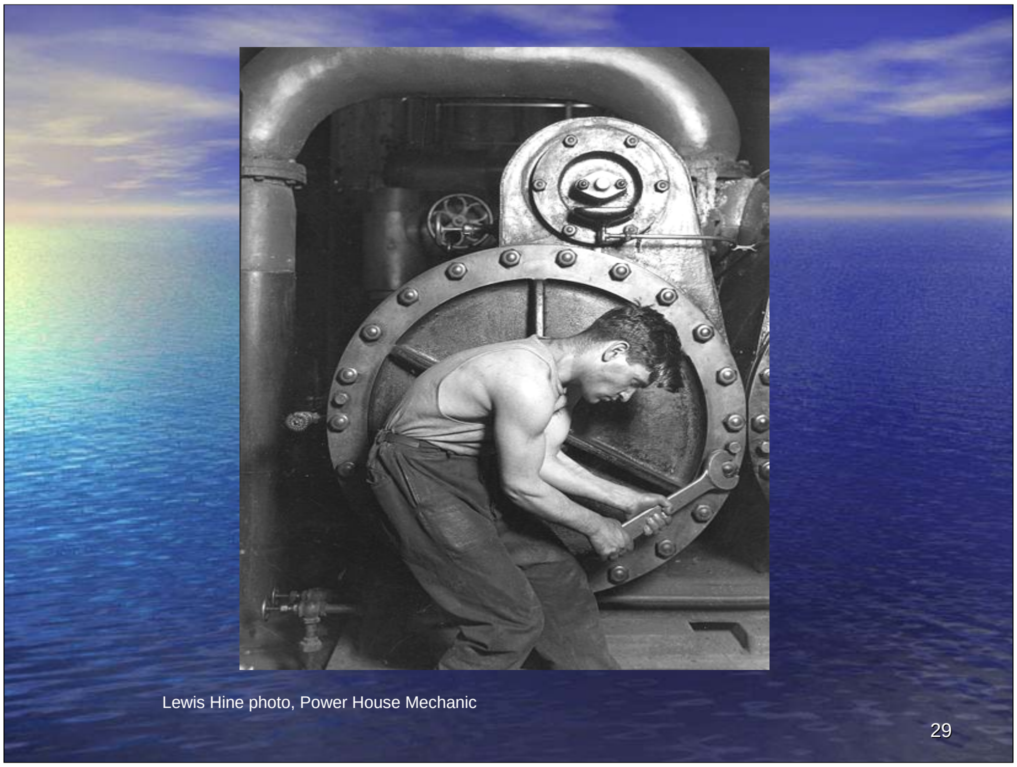
Use of Other Statistical Methods and Software, e.g., ProUCL®

- The Unified Guidance provides a reasonable suite of methods, but by no means exhaustive
- Statistical literature references to other possible tests are provided
- The guidance suggests use of R-script and ProUCL for certain applications. Many other commercial and proprietary software may be available.

The intention of the Unified Guidance is to provide a reasonable set of methods and procedures which can address most RCRA groundwater monitoring and testing concerns. However, the guidance is also clear that it is not intended to be exhaustive. Considerable statistical literature on methods and procedures is available, and could certainly be considered. A regulatory agency will still need to evaluate and approve such alternatives in light of the general RCRA performance criteria.

In this guidance, we have provided certain uses of R, a public source software, for difficult and tedious calculations. The R-library is considerable, and users may wish to consider other applications. We also referenced ProUCL as an option for identifying confidence limits for a lognormal distribution. It also contains distributional fitting for the gamma distribution, which may prove superior in some cases. Gibbons and Bhaumik have also developed papers and techniques on applying the gamma distribution.

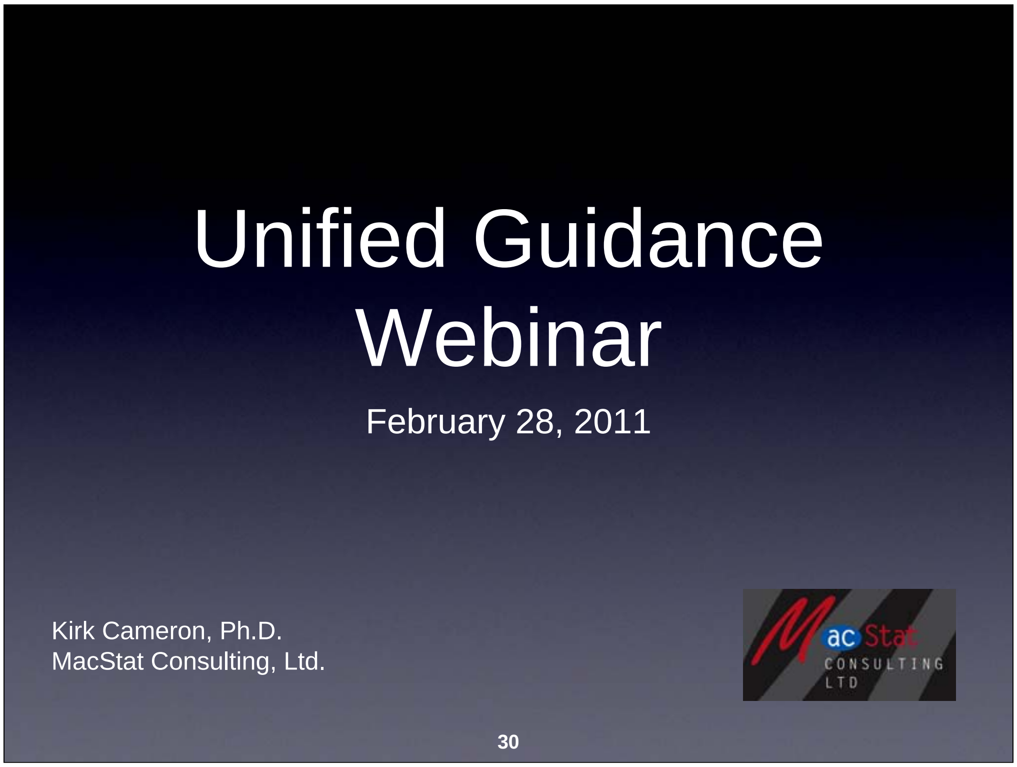
Facilities and regulators will undoubtedly need to consider other commercial and proprietary software as well. It was not possible to do such evaluations within the limited scope of the present guidance. Hopefully, this guidance offers some additional tools which might be used to assess the general RCRA performance criteria.



Lewis Hine photo, Power House Mechanic

29

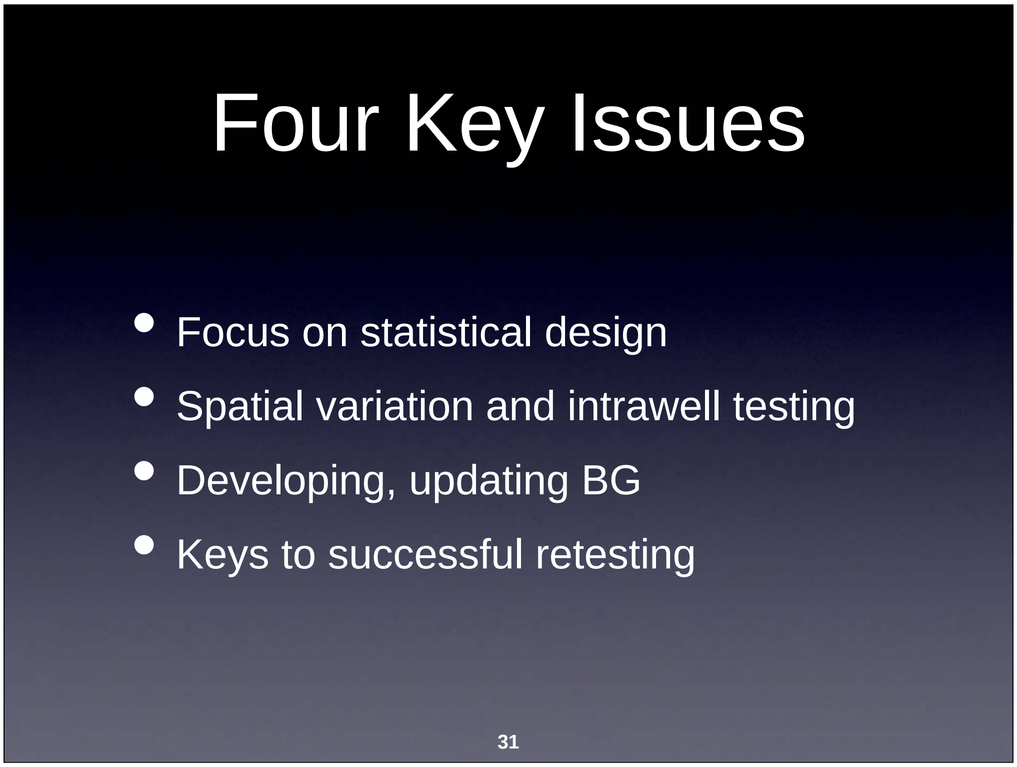
In conclusion, once you have and know the tools, applying the guidance will be much more understandable. Dr. Kirk Cameron will now continue with his presentation on some major statistical “nuts and bolts” found in the guidance.



Unified Guidance Webinar

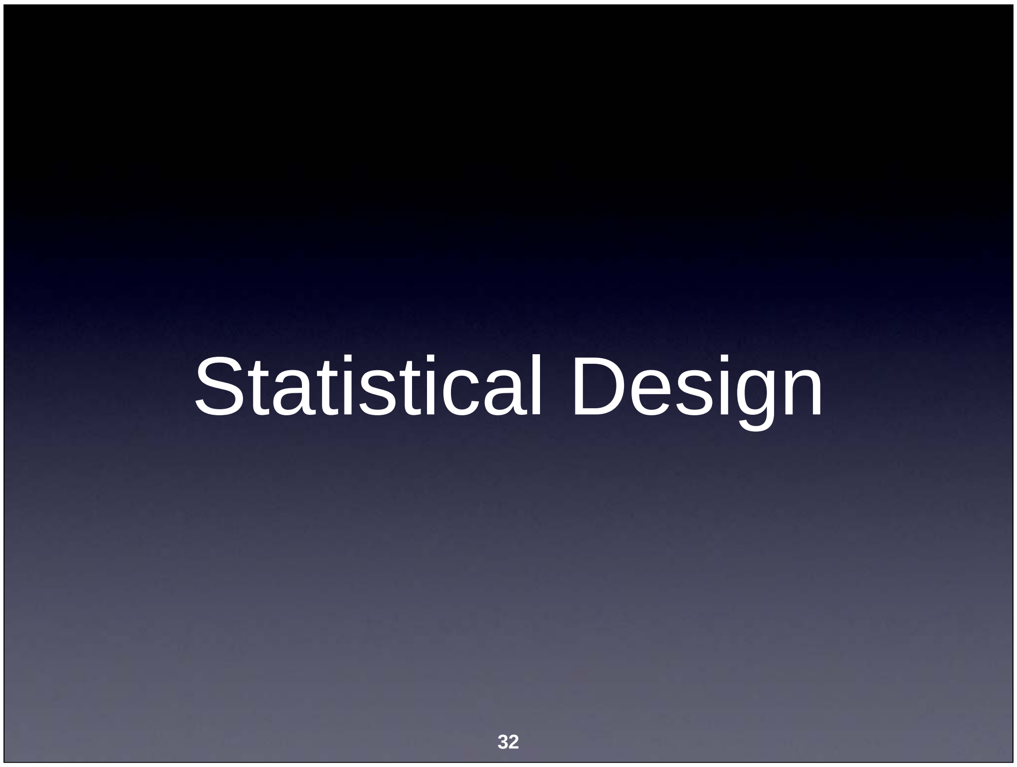
February 28, 2011

Kirk Cameron, Ph.D.
MacStat Consulting, Ltd.



Four Key Issues

- Focus on statistical design
- Spatial variation and intrawell testing
- Developing, updating BG
- Keys to successful retesting

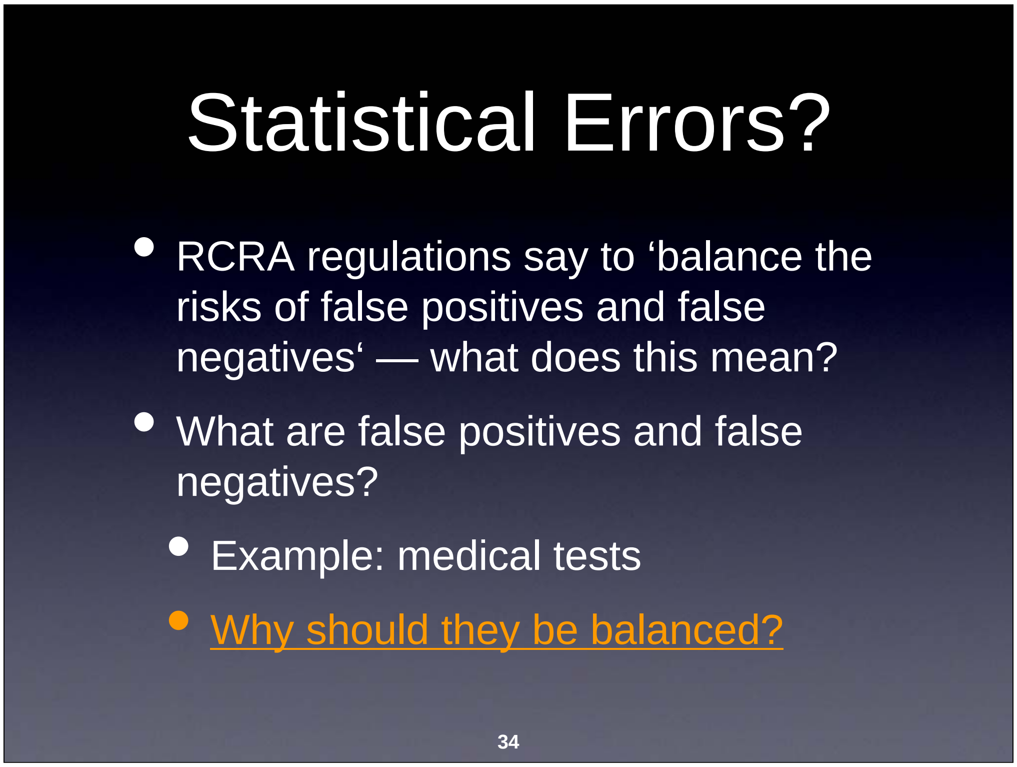


Statistical Design

32

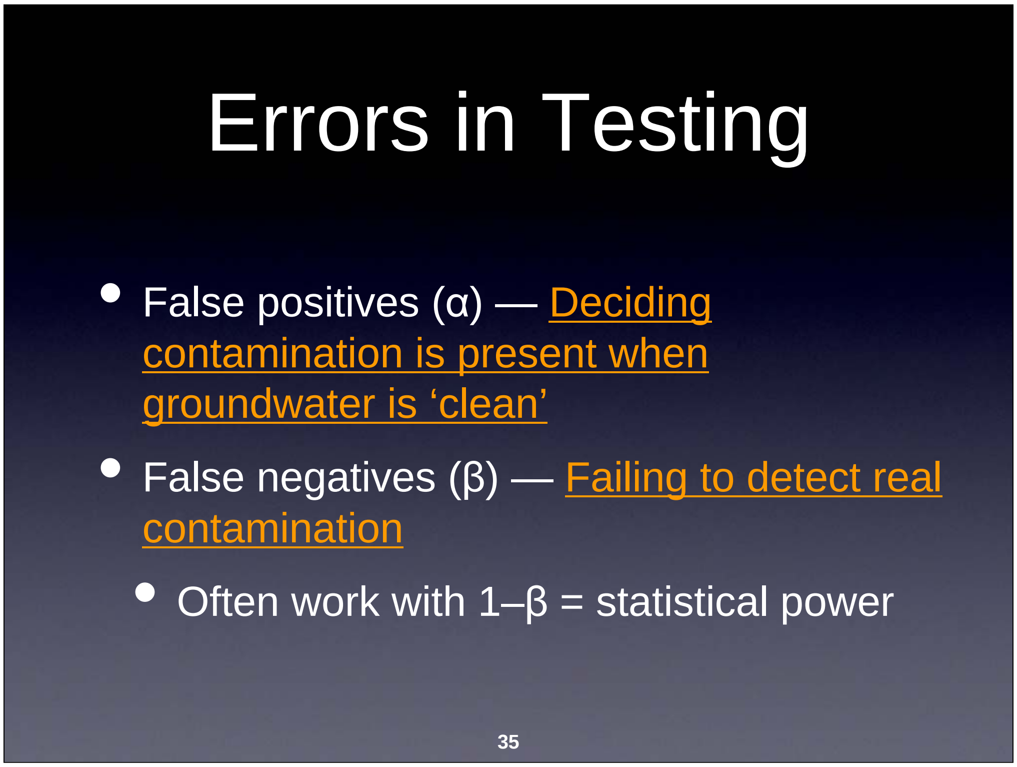
Designed for Good

- UG promotes good statistical design principles
  - Do it up front
  - Refine over life of facility



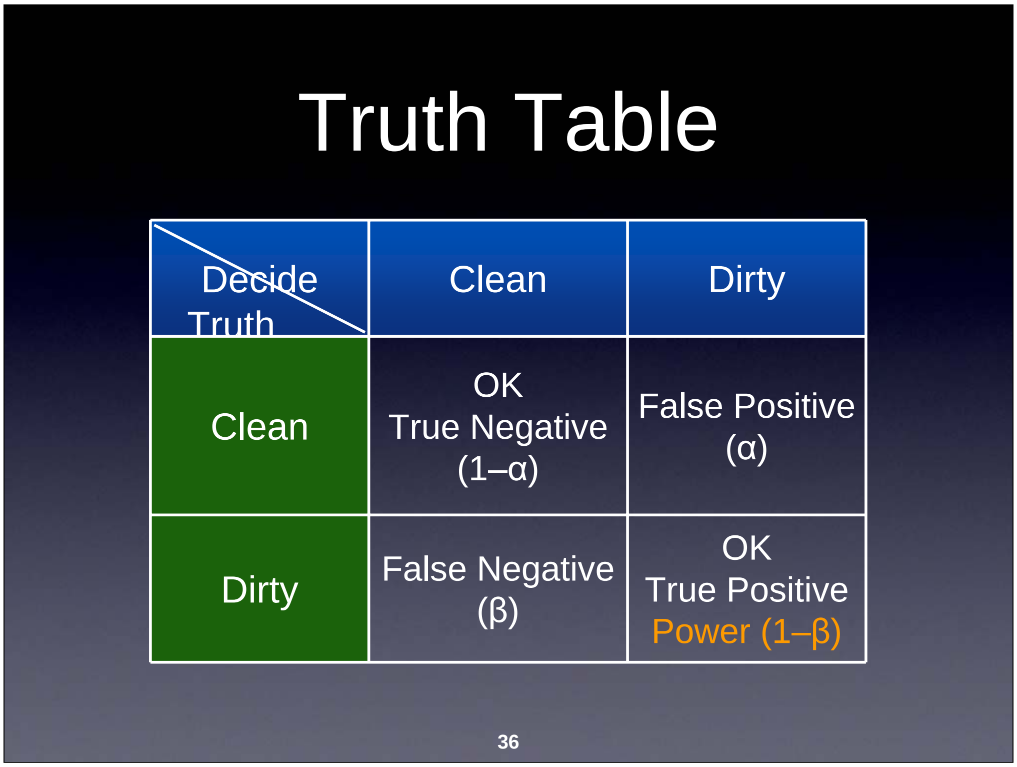
Statistical Errors?

- RCRA regulations say to ‘balance the risks of false positives and false negatives‘ — what does this mean?
- What are false positives and false negatives?
  - Example: medical tests
  - Why should they be balanced?



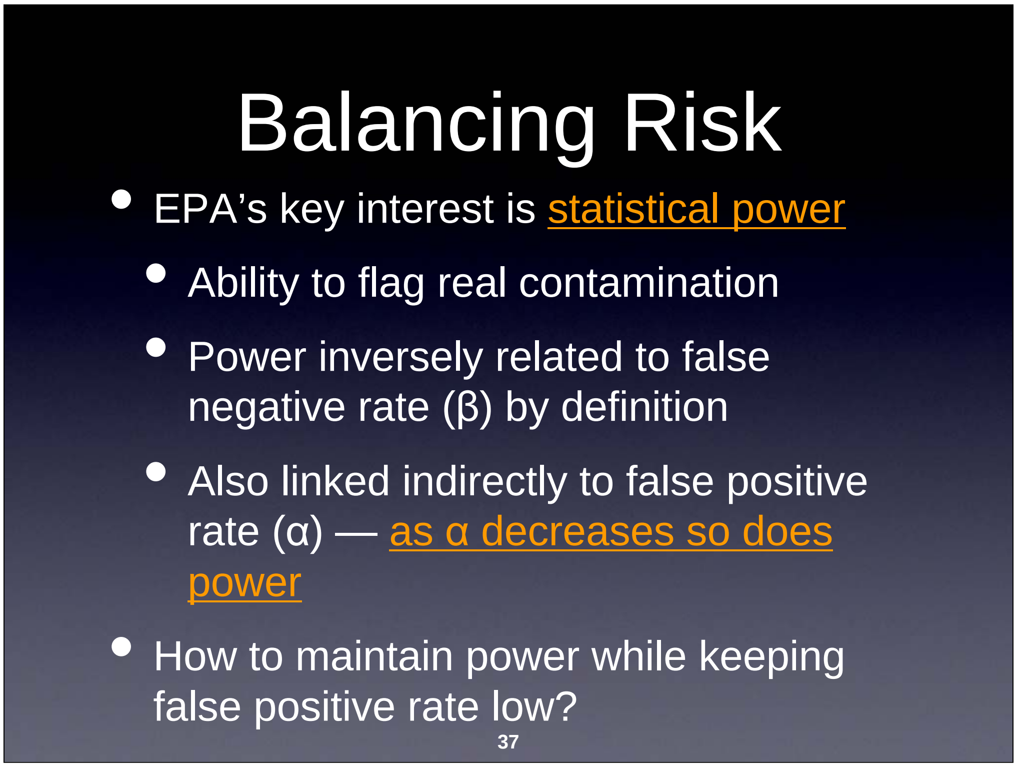
Errors in Testing

- False positives (α) — Deciding contamination is present when groundwater is ‘clean’
- False negatives (β) — Failing to detect real contamination
    - Often work with 1–β = statistical power



Truth Table

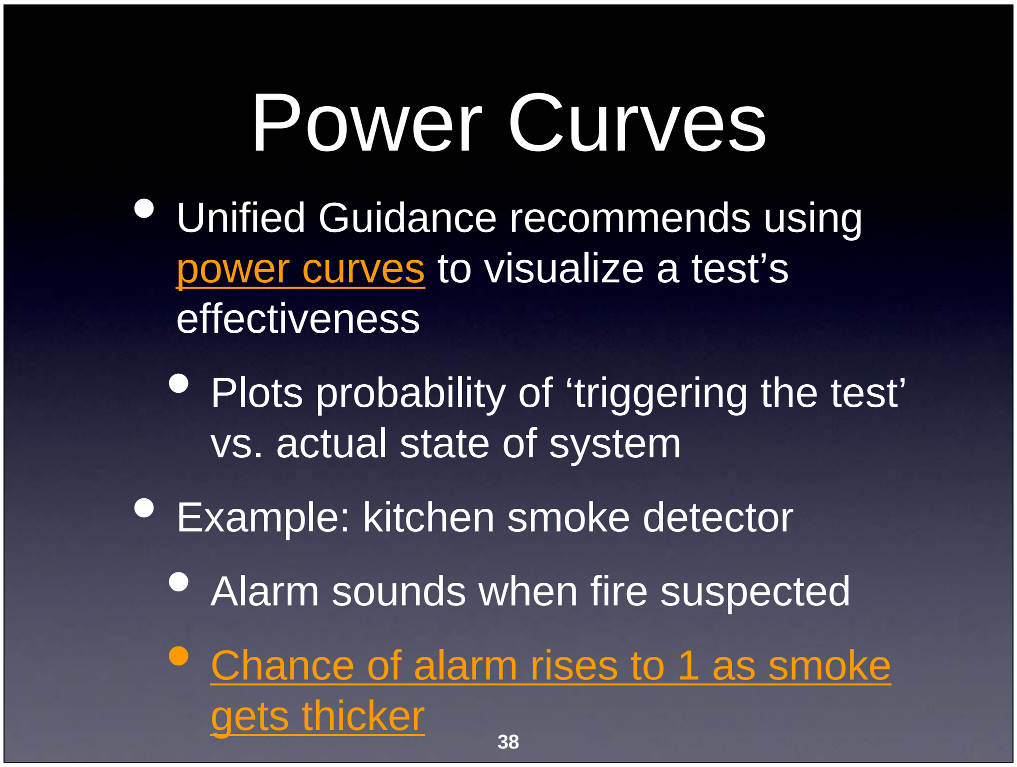
Decide Truth	Clean	Dirty
Clean	OK True Negative (1–α)	False Positive (α)
Dirty	False Negative (β)	OK True Positive Power (1–β)



Balancing Risk

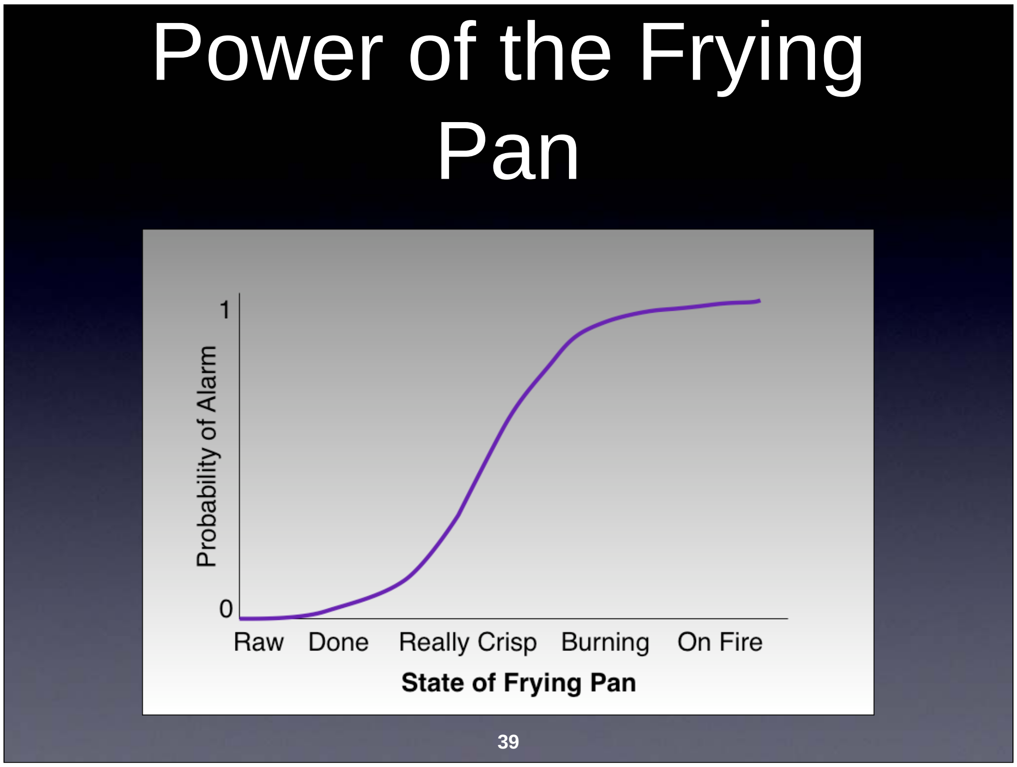
- EPA’s key interest is statistical power
  - Ability to flag real contamination
  - Power inversely related to false negative rate (β) by definition
  - Also linked indirectly to false positive rate (α) — as α decreases so does power
- How to maintain power while keeping false positive rate low?

37

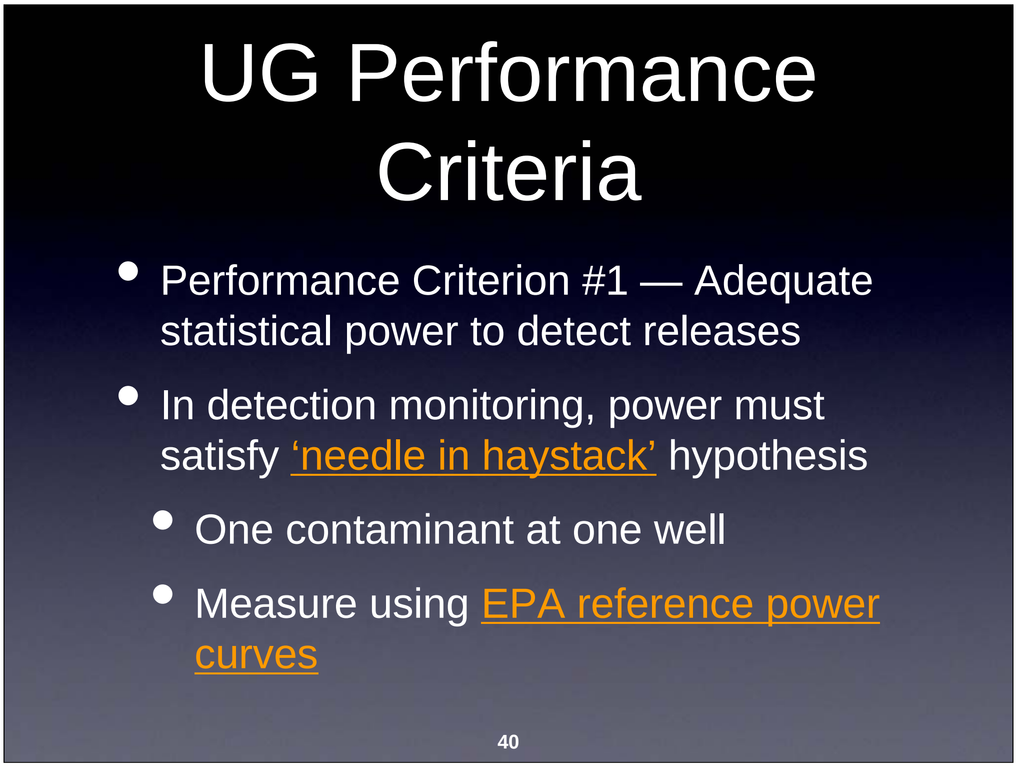


Power Curves

- Unified Guidance recommends using power curves to visualize a test’s effectiveness
  - Plots probability of ‘triggering the test’ vs. actual state of system
- Example: kitchen smoke detector
  - Alarm sounds when fire suspected
  - Chance of alarm rises to 1 as smoke gets thicker



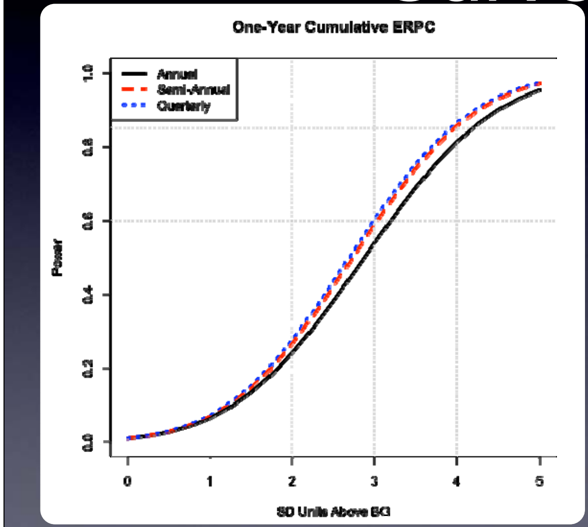
Power of the Frying Pan



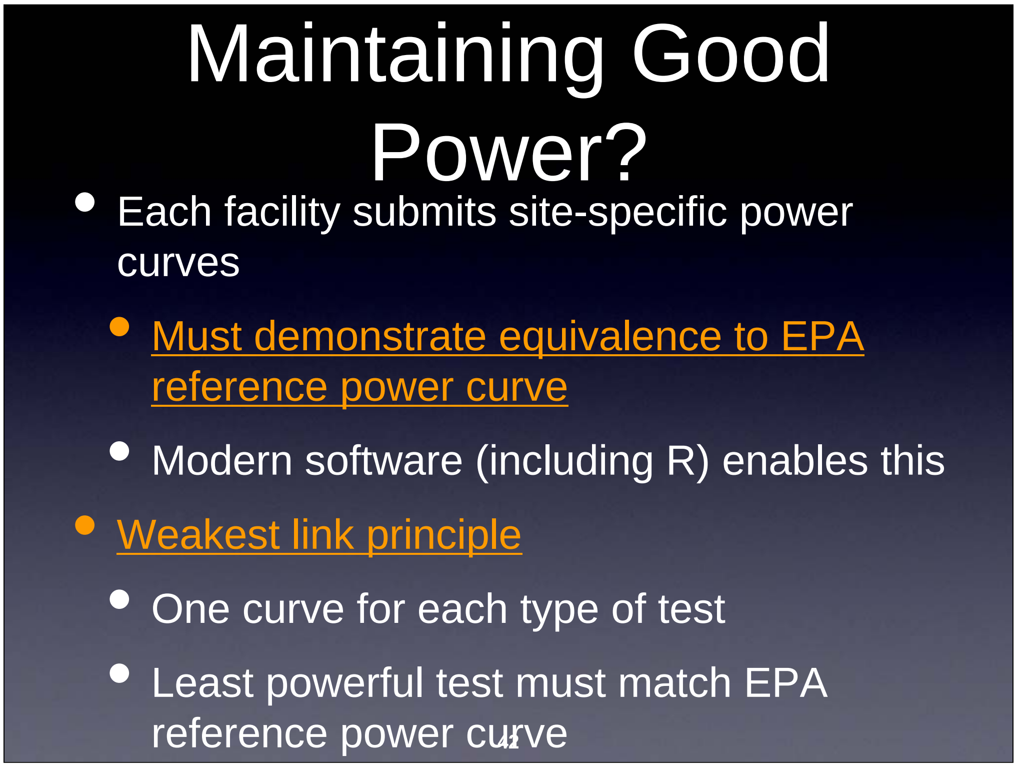
UG Performance Criteria

- Performance Criterion #1 — Adequate statistical power to detect releases
- In detection monitoring, power must satisfy ‘needle in haystack’ hypothesis
  - One contaminant at one well
  - Measure using EPA reference power curves

Reference Power Curves

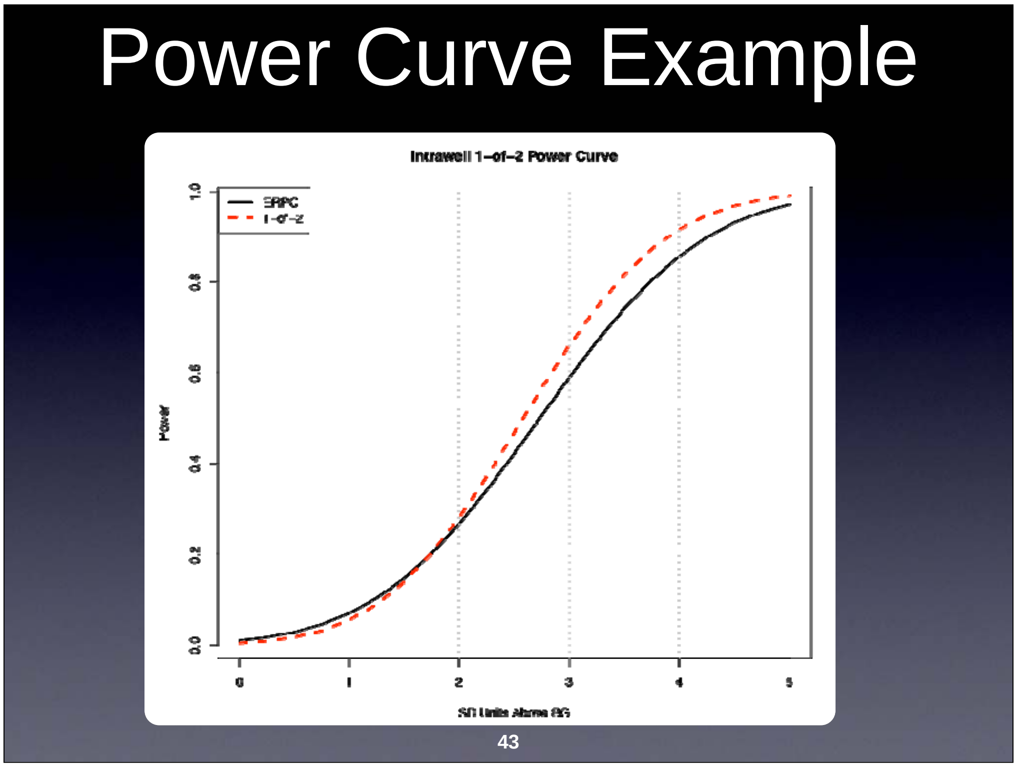


- Users pick curve based on evaluation frequency
  - Annual, semi-annual, quarterly
- Key targets: 55-60% at 3 SDs, 80-85% at 4 SDs



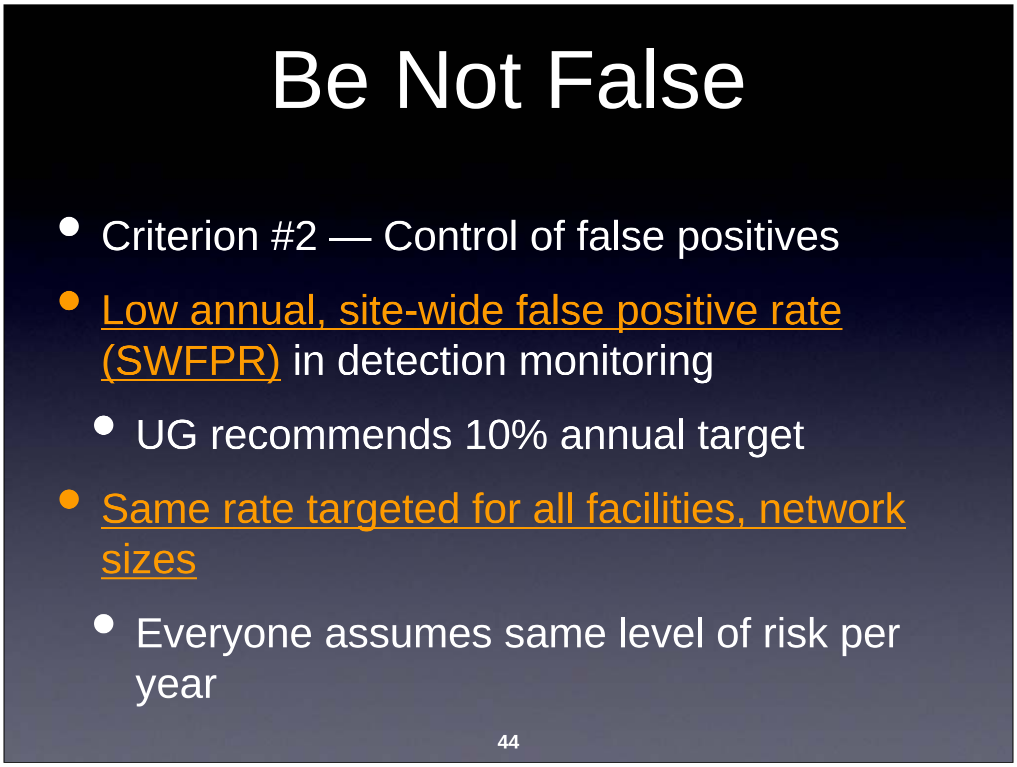
Maintaining Good Power?

- Each facility submits site-specific power curves
  - Must demonstrate equivalence to EPA reference power curve
  - Modern software (including R) enables this
- Weakest link principle
  - One curve for each type of test
  - Least powerful test must match EPA reference power curve



Power Curve Example

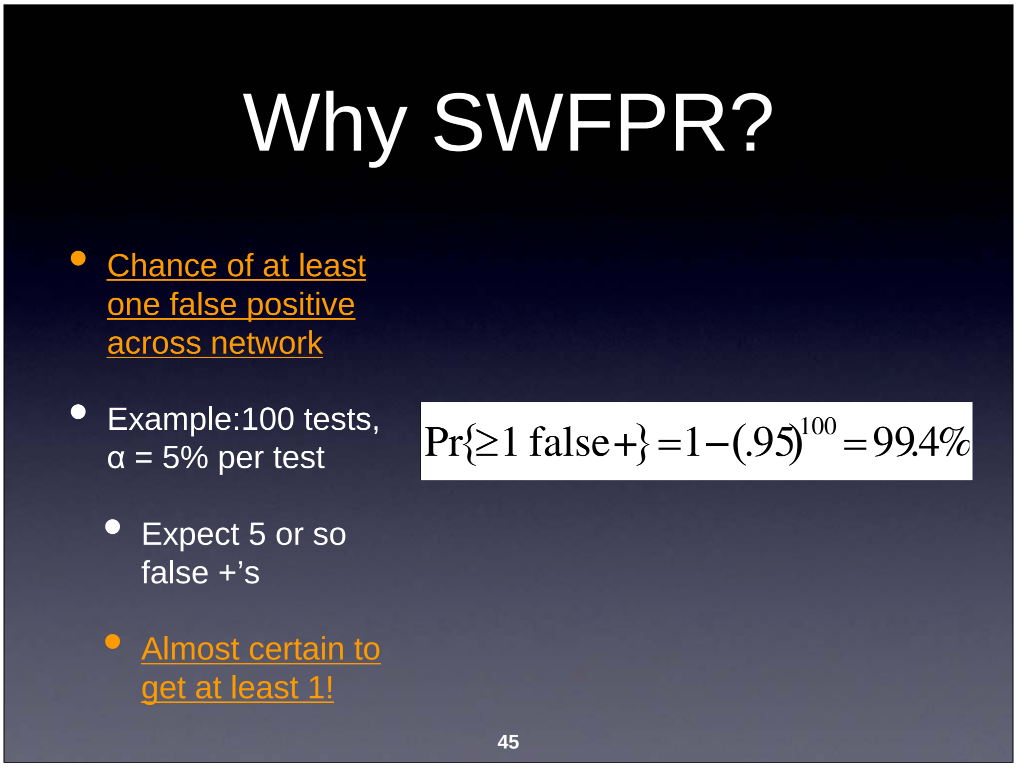
43



Be Not False

- Criterion #2 — Control of false positives
- Low annual, site-wide false positive rate (SWFPR) in detection monitoring
  - UG recommends 10% annual target
- Same rate targeted for all facilities, network sizes
  - Everyone assumes same level of risk per year

44

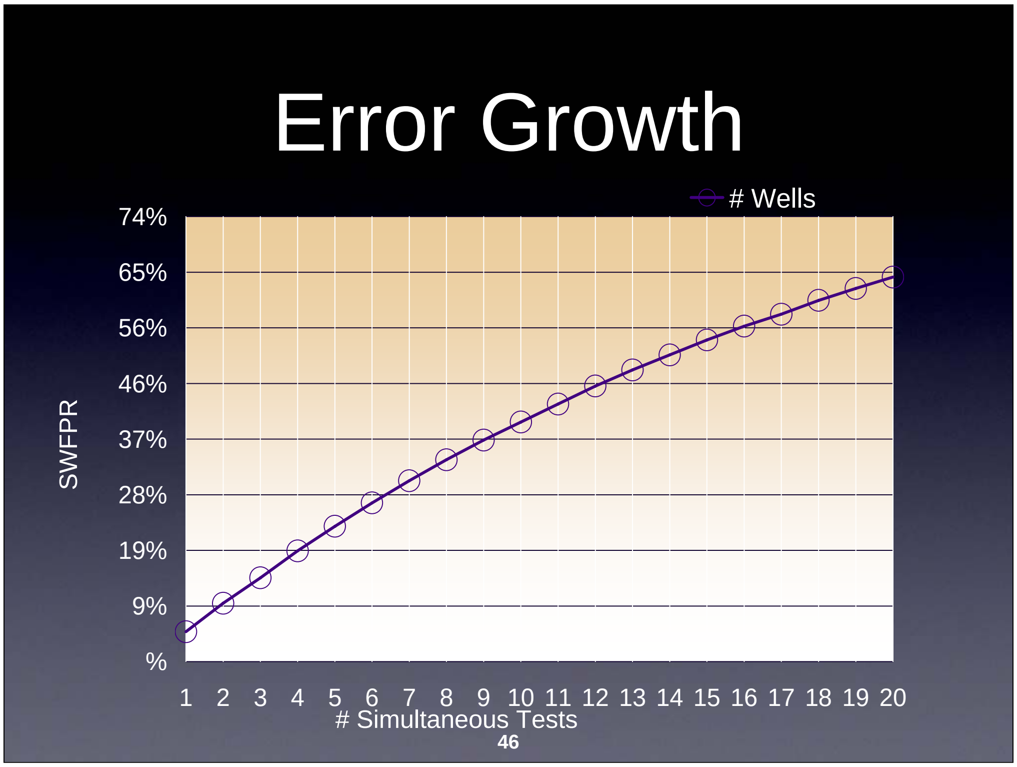


Why SWFPR?

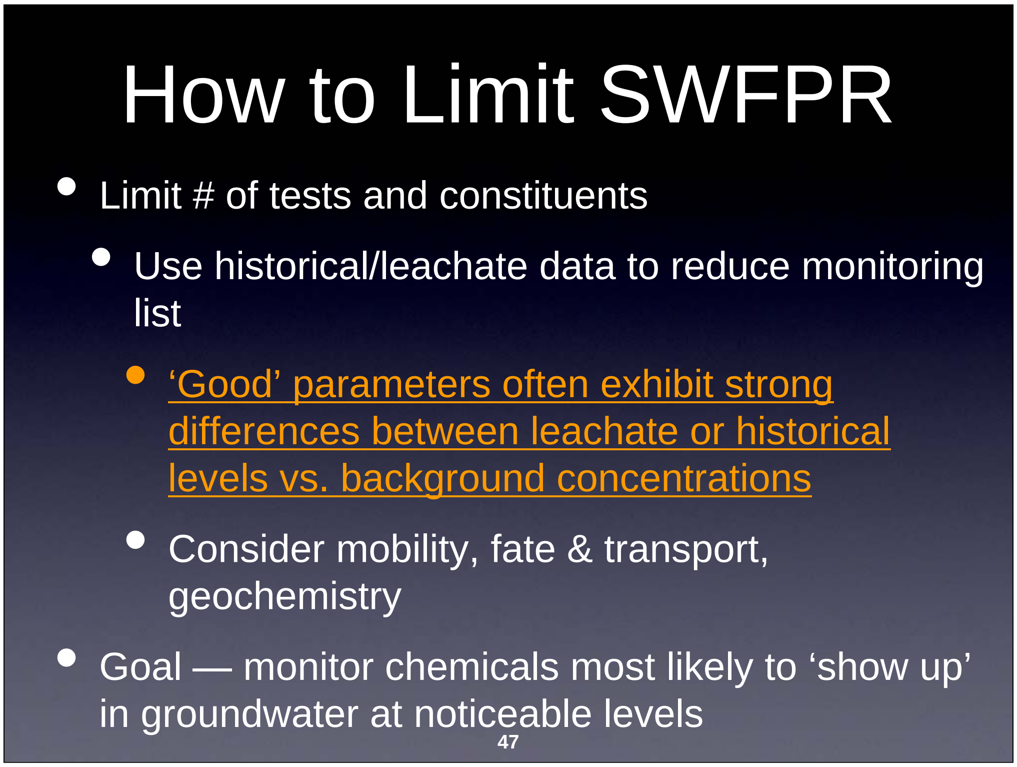
- Chance of at least one false positive across network
- Example:100 tests, α = 5% per test
  - Expect 5 or so false +’s
  - Almost certain to get at least 1!

$$\Pr\{\geq 1 \text{ false}+\} = 1 - (.95)^{100} = 99.4\%$$

45

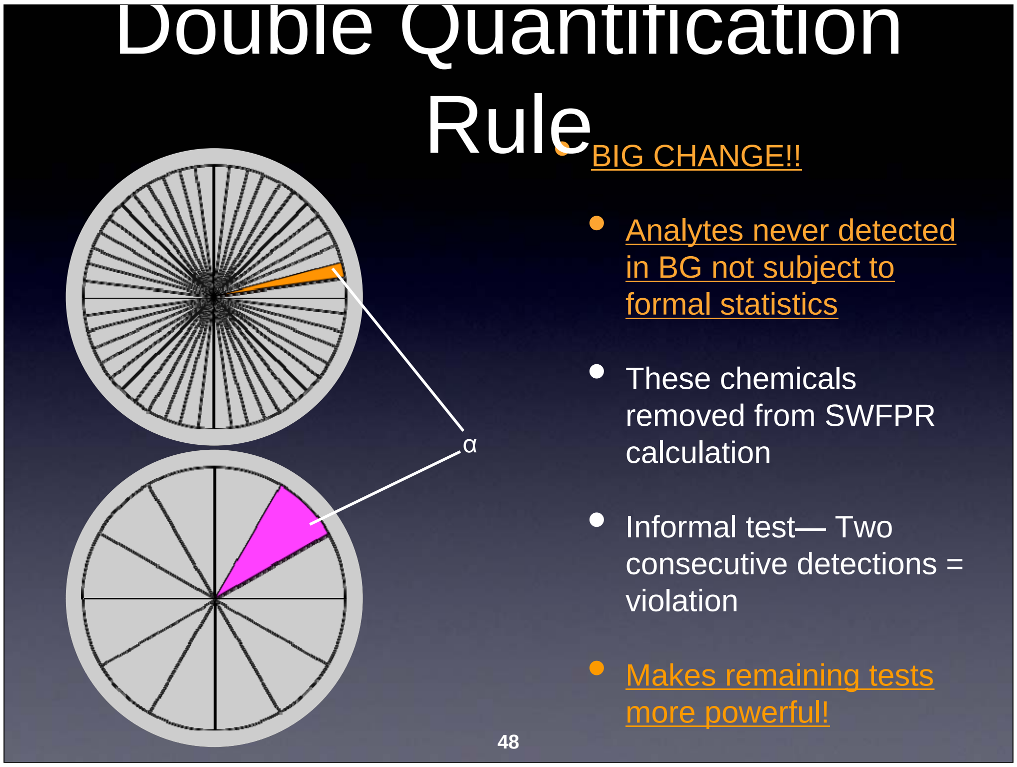


Error Growth



How to Limit SWFPR

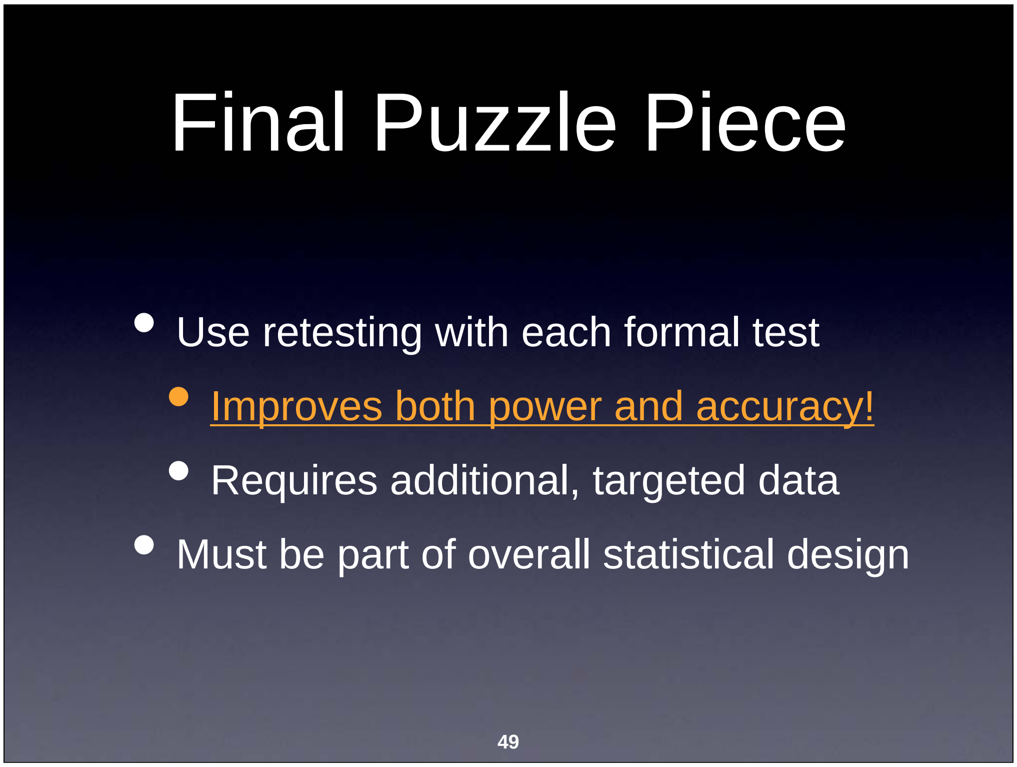
- Limit # of tests and constituents
  - Use historical/leachate data to reduce monitoring list
    - ‘Good’ parameters often exhibit strong differences between leachate or historical levels vs. background concentrations
    - Consider mobility, fate & transport, geochemistry
- Goal — monitor chemicals most likely to ‘show up’ in groundwater at noticeable levels



Double Quantification Rule

- BIG CHANGE!!
  - Analytes never detected in BG not subject to formal statistics
  - These chemicals removed from SWFPR calculation
  - Informal test— Two consecutive detections = violation
  - Makes remaining tests more powerful!

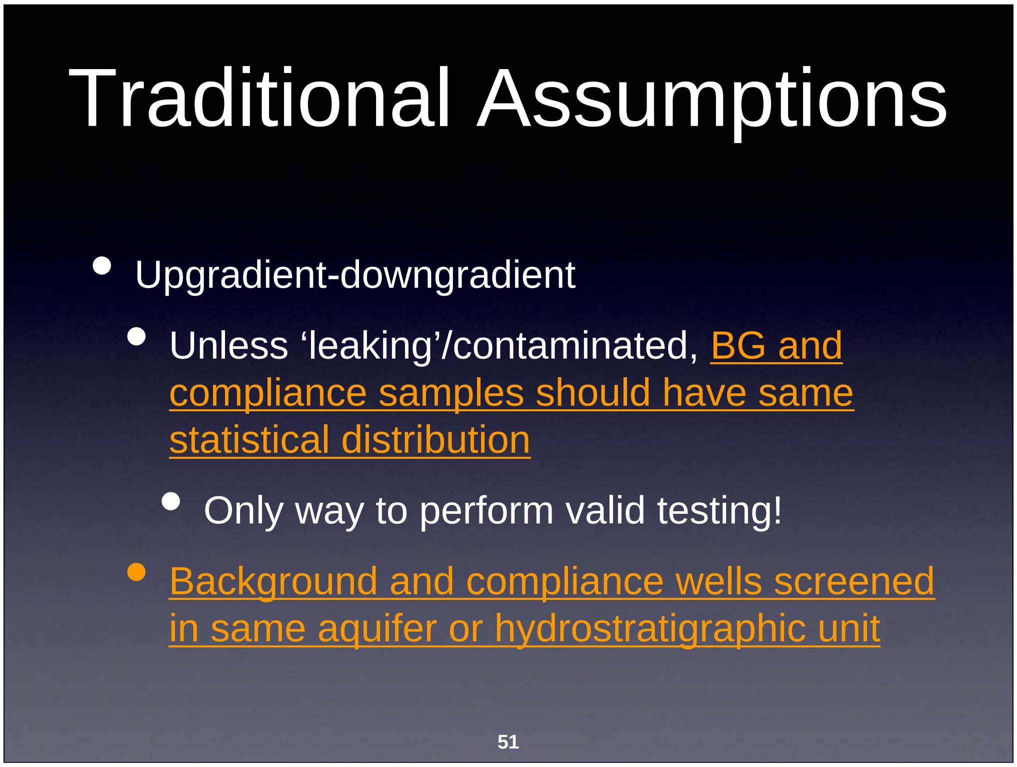
Both ‘pizzas’ in this graphic are designed to represent a fixed annual false positive rate target of 10% across the site as a whole; the upper pizza illustrates the difficulty of doing statistical analysis on a large number of wells and constituents: to keep the SWFPR at 10%, the individual test α must be set quite low, leading to a less powerful test; in the lower pizza, constituents never detected in on-site background have been removed from formal statistical analysis; the smaller number of remaining tests can be run at a higher α and consequently higher power, thus improving the statistical analysis on those constituents that have actually been observed on-site.



Final Puzzle Piece

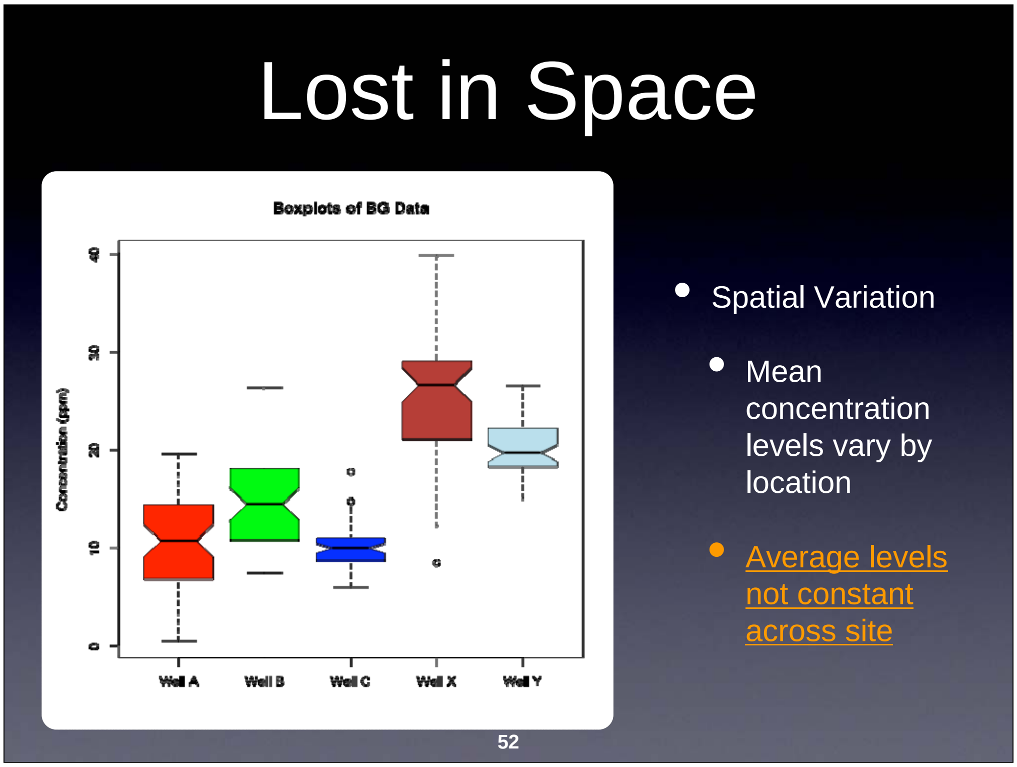
- Use retesting with each formal test
  - Improves both power and accuracy!
  - Requires additional, targeted data
- Must be part of overall statistical design

Spatial Variation, Intrawell Testing



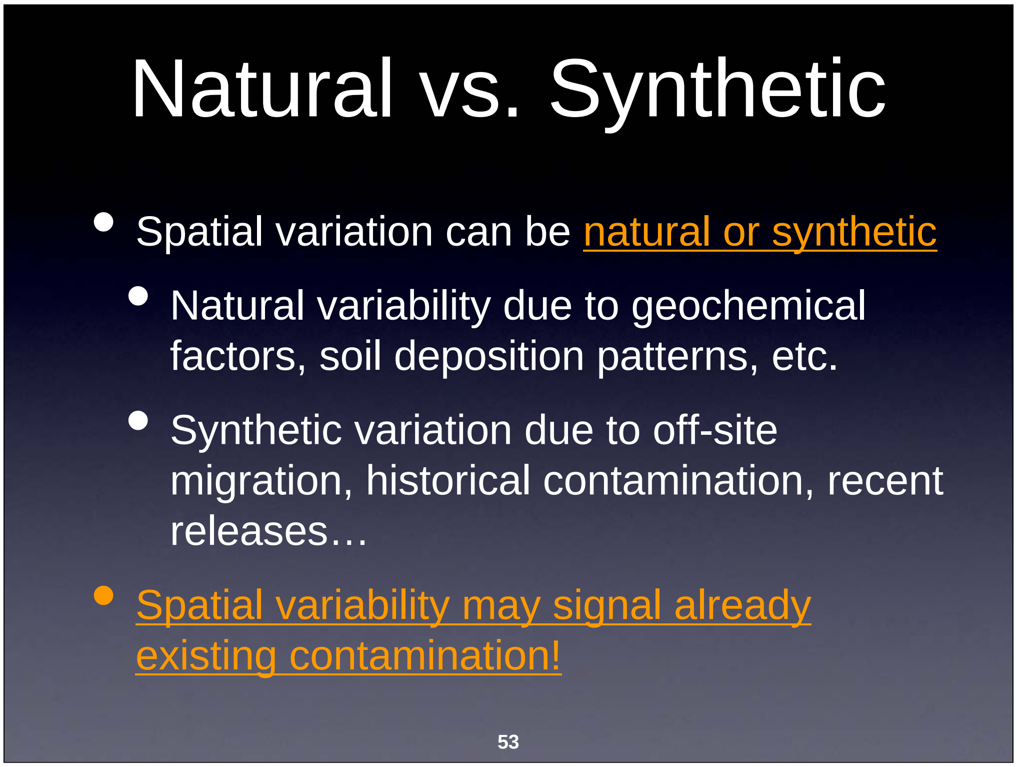
Traditional Assumptions

- Upgradient-downgradient
  - Unless ‘leaking’/contaminated, BG and compliance samples should have same statistical distribution
    - Only way to perform valid testing!
  - Background and compliance wells screened in same aquifer or hydrostratigraphic unit



Lost in Space

- Spatial Variation
  - Mean concentration levels vary by location
  - Average levels not constant across site

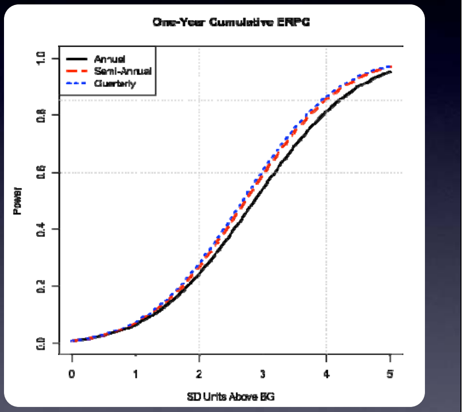


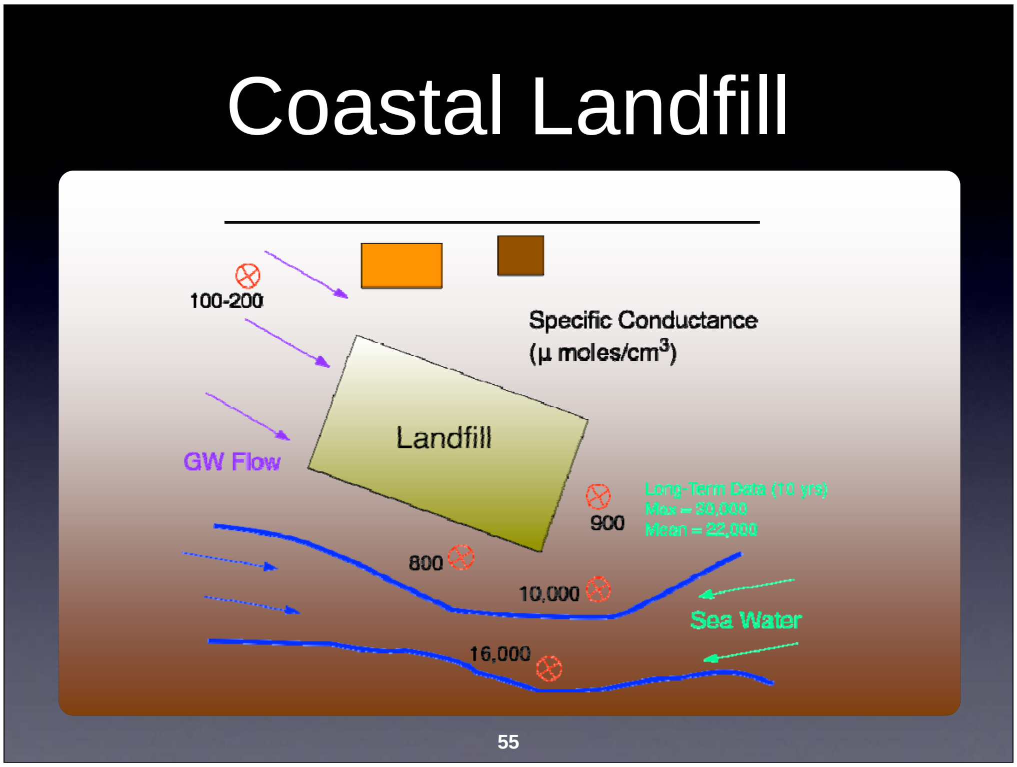
Natural vs. Synthetic

- Spatial variation can be natural or synthetic
  - Natural variability due to geochemical factors, soil deposition patterns, etc.
  - Synthetic variation due to off-site migration, historical contamination, recent releases…
- Spatial variability may signal already existing contamination!

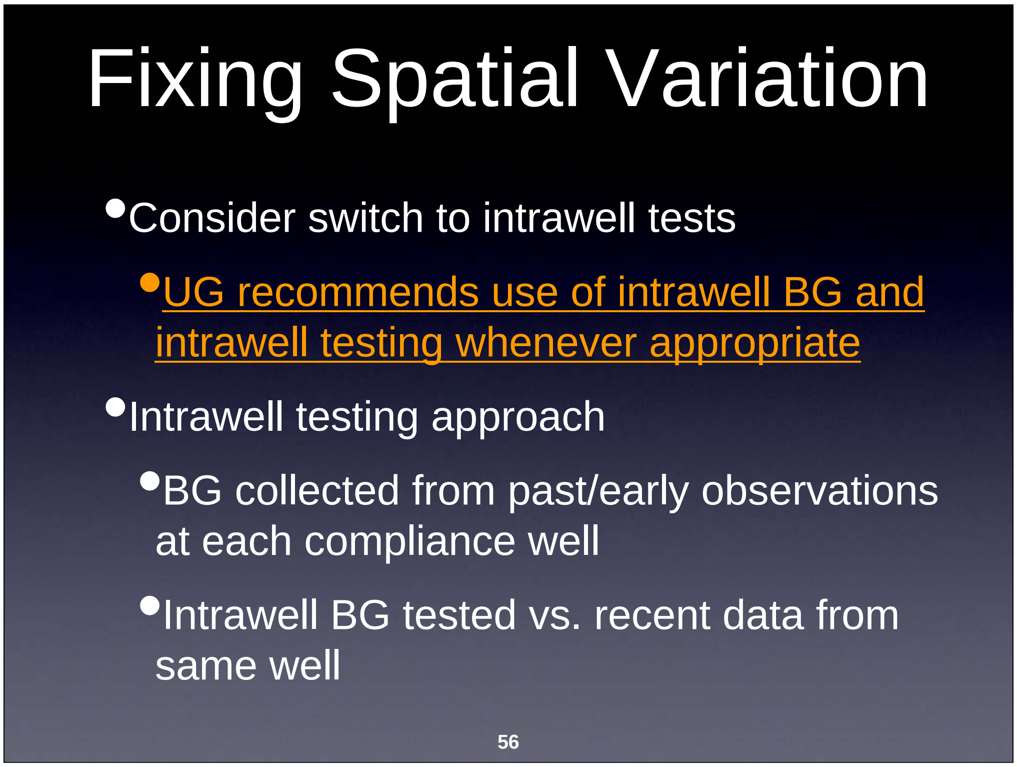
Impact of Spatial Variation

- Statistical test answers wrong question!
  - Can’t compare apples-to-apples
  - Example— upgradient-downgradient test
    - Suppose sodium values naturally 20 ppm (4 SDs) higher than background on average?
    - 80%+ power essentially meaningless!



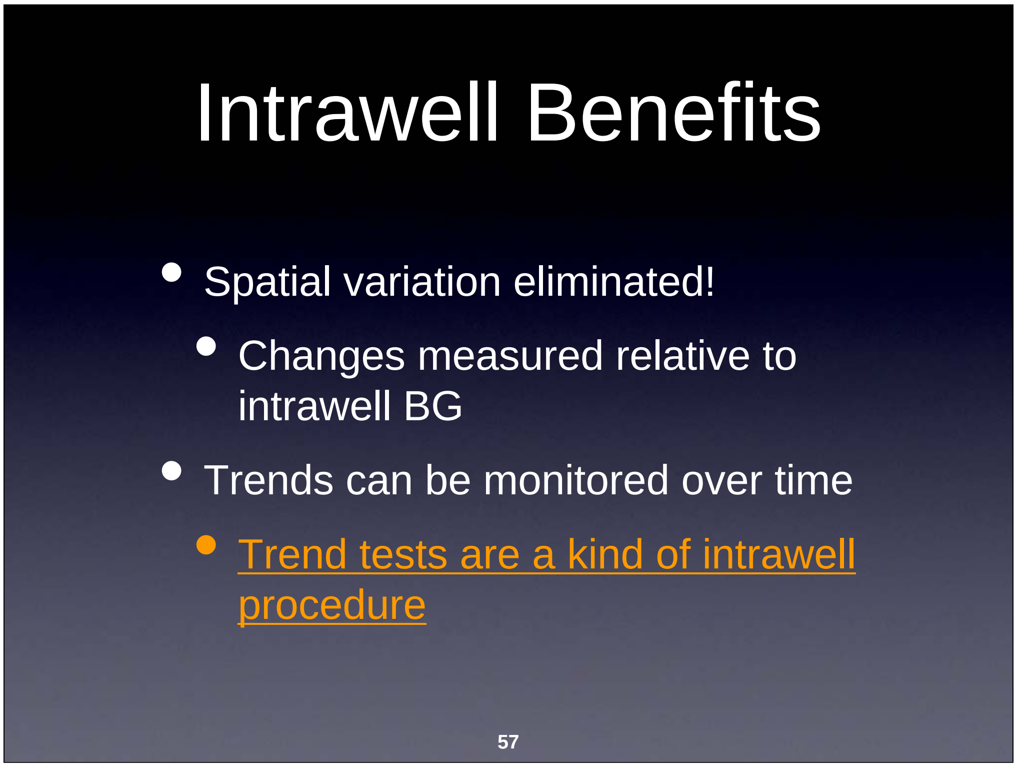


Coastal Landfill



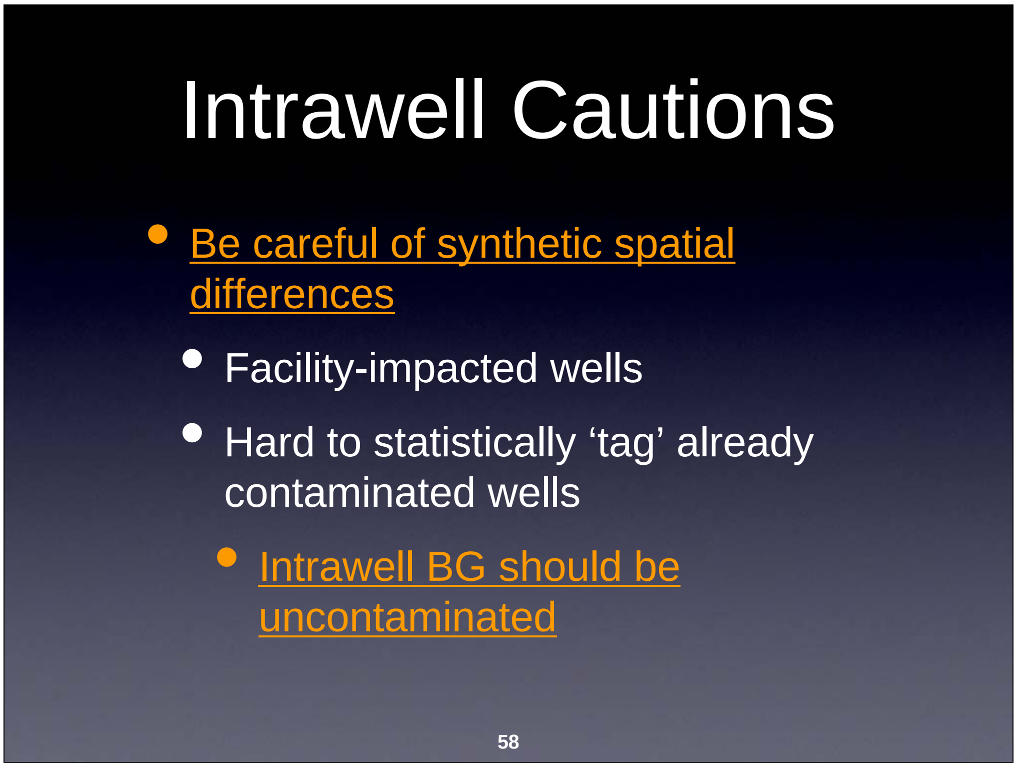
Fixing Spatial Variation

- Consider switch to intrawell tests
  - UG recommends use of intrawell BG and intrawell testing whenever appropriate
- Intrawell testing approach
  - BG collected from past/early observations at each compliance well
  - Intrawell BG tested vs. recent data from same well



Intrawell Benefits

- Spatial variation eliminated!
  - Changes measured relative to intrawell BG
- Trends can be monitored over time
  - Trend tests are a kind of intrawell procedure

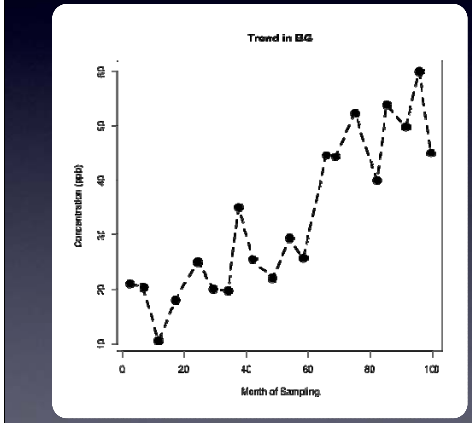


Intrawell Cautions

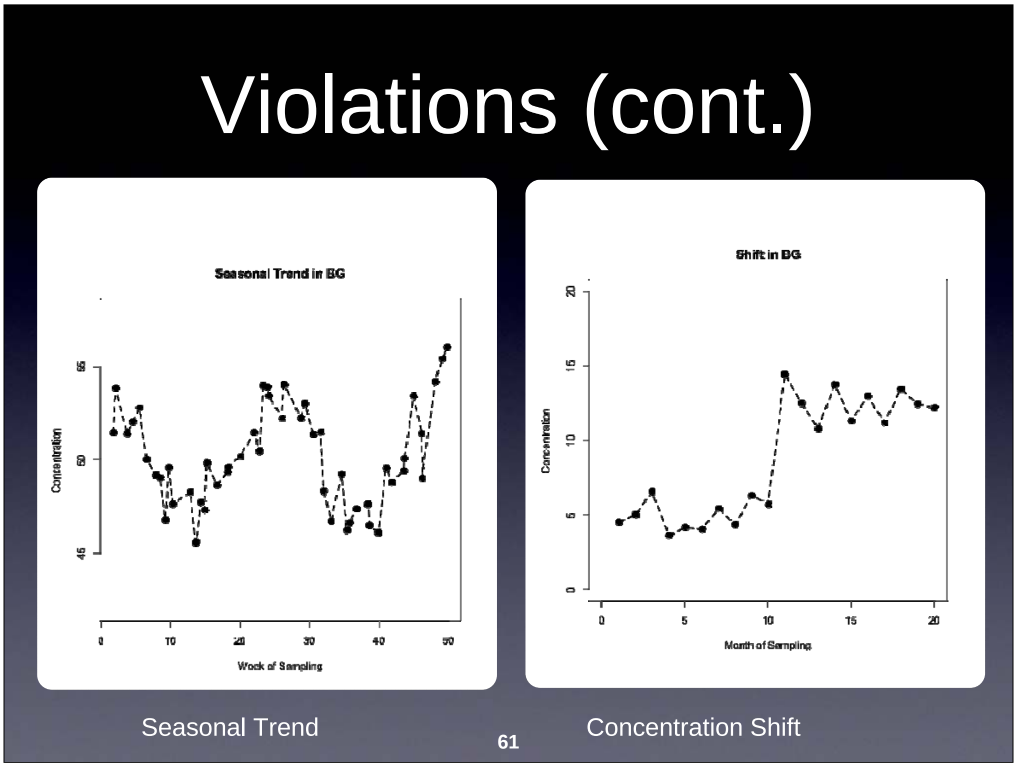
- Be careful of synthetic spatial differences
  - Facility-impacted wells
  - Hard to statistically ‘tag’ already contaminated wells
    - Intrawell BG should be uncontaminated

Developing, Updating Background

BG Assumptions



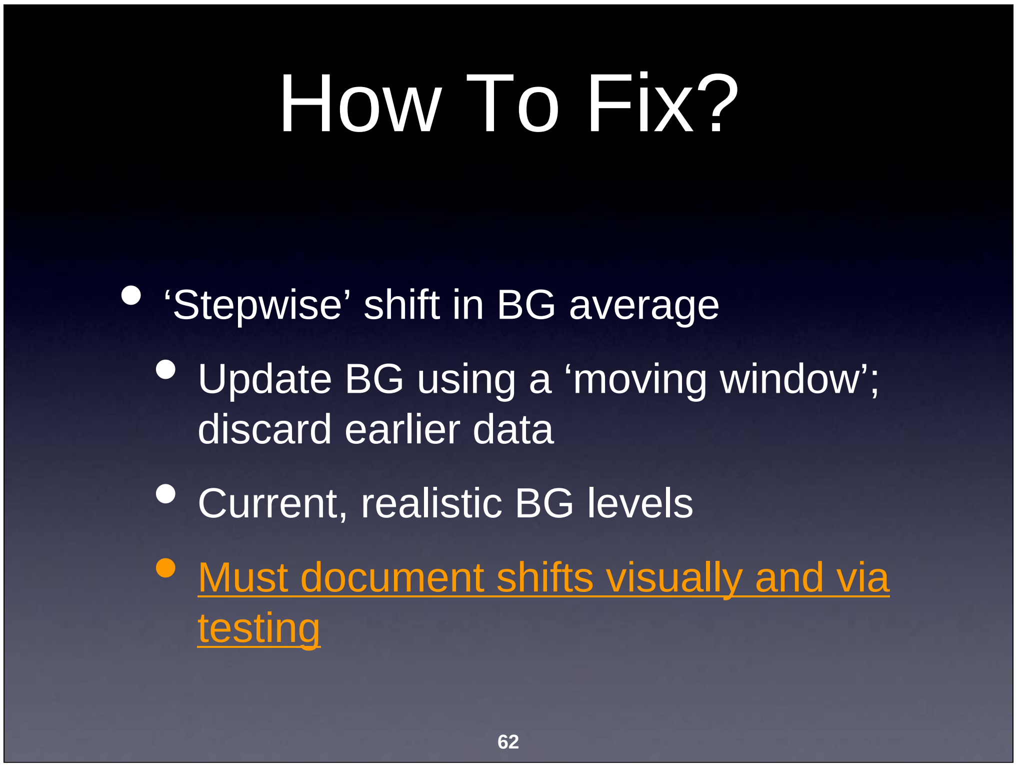
- Levels should be stable (stationary) over time
- Look for violations
  - Distribution of BG concentrations changing
  - Trend, shift, or cyclical pattern evident



Violations (cont.)

Seasonal Trend

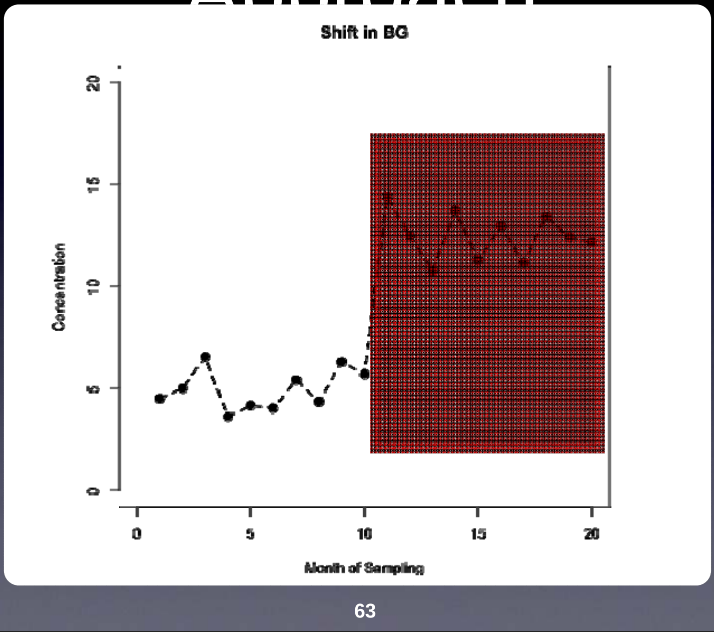
Concentration Shift

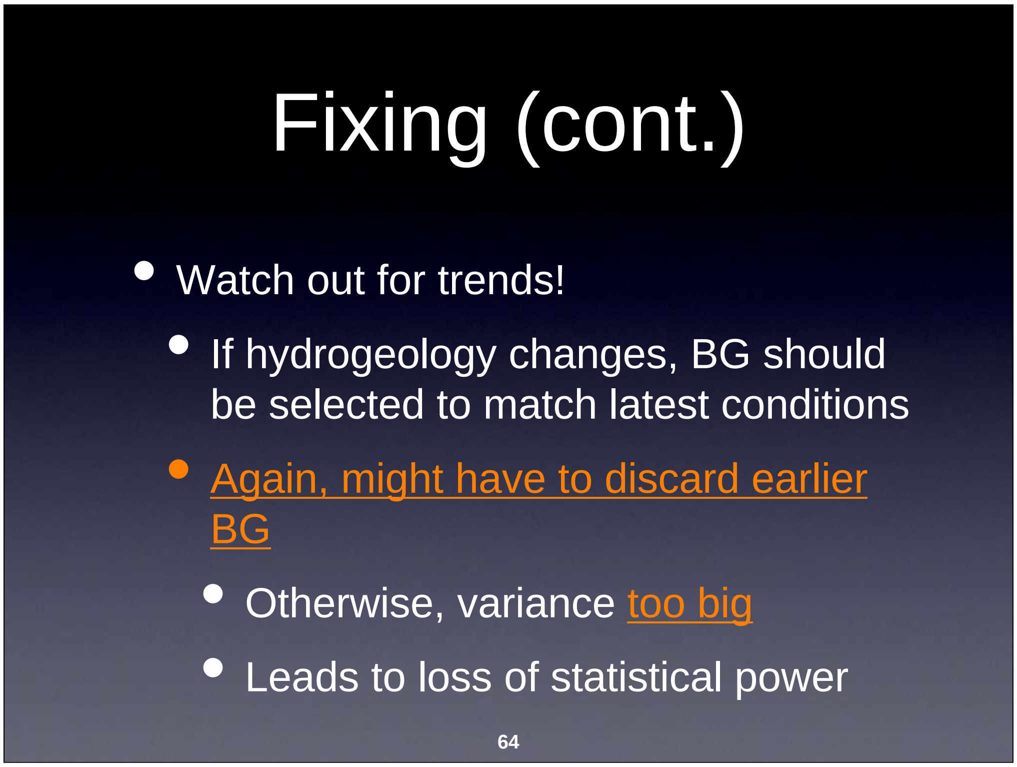


How To Fix?

- ‘Stepwise’ shift in BG average
  - Update BG using a ‘moving window’; discard earlier data
  - Current, realistic BG levels
  - Must document shifts visually and via testing

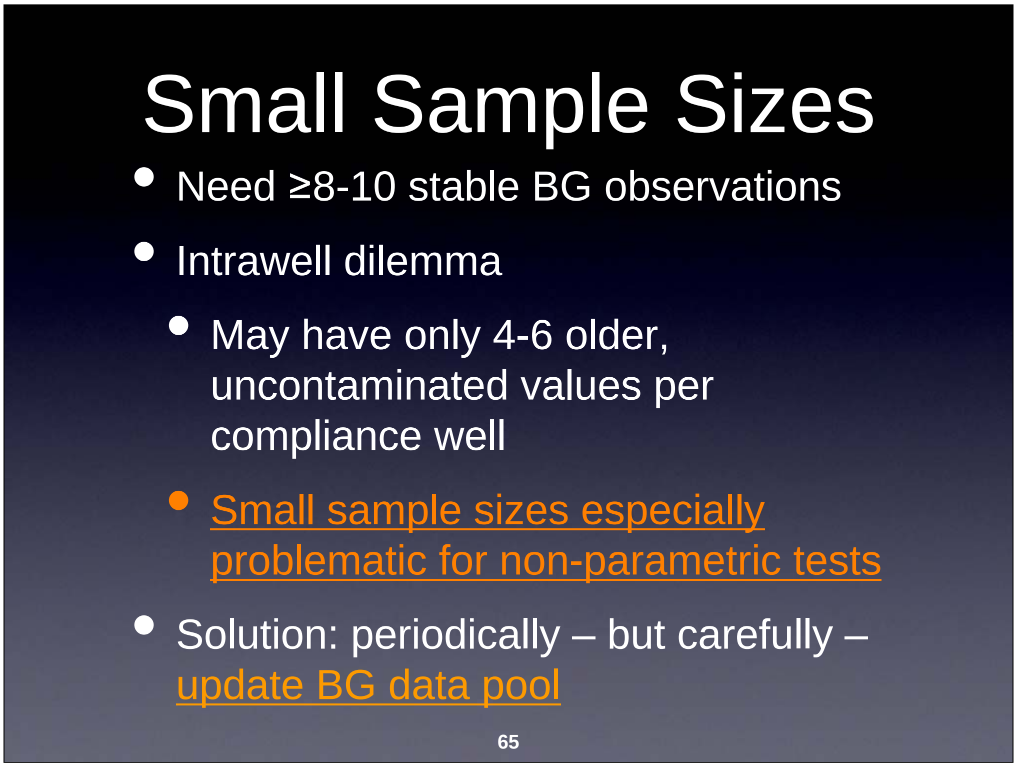
Moving Window Approach





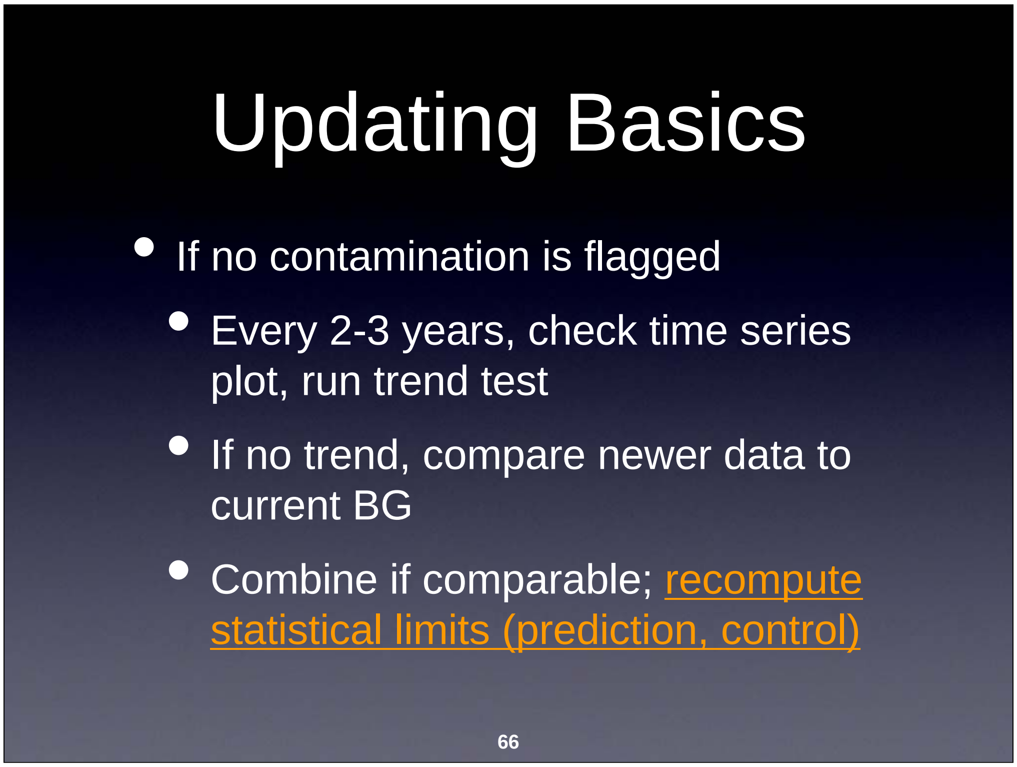
Fixing (cont.)

- Watch out for trends!
  - If hydrogeology changes, BG should be selected to match latest conditions
  - Again, might have to discard earlier BG
    - Otherwise, variance too big
    - Leads to loss of statistical power



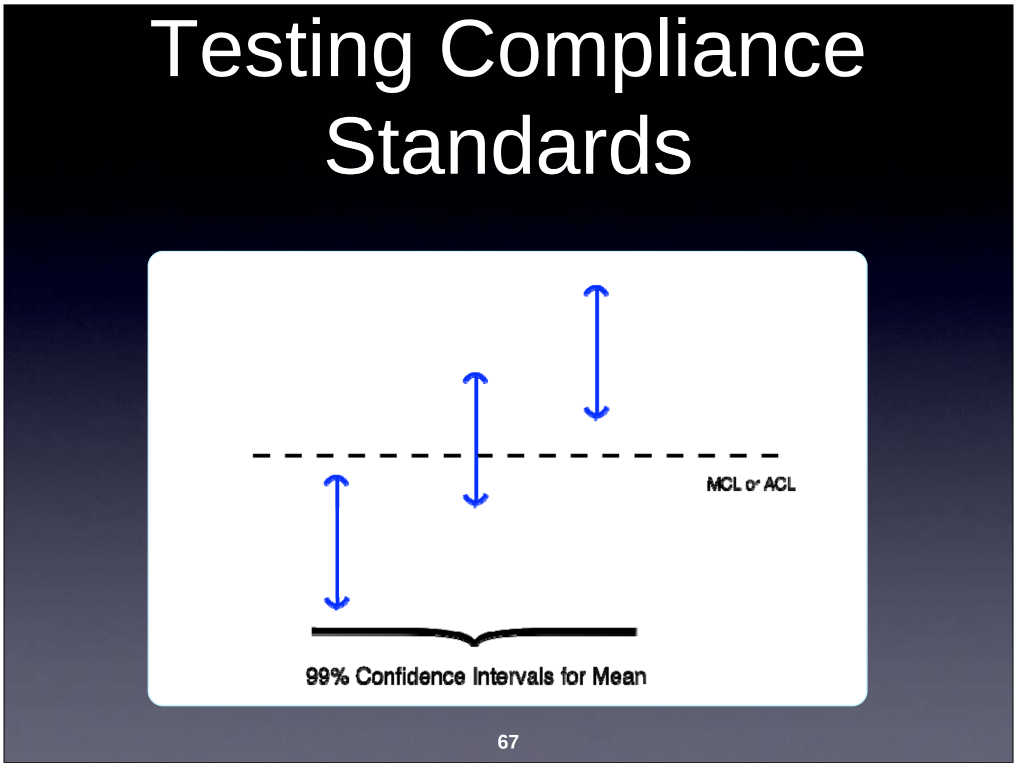
Small Sample Sizes

- Need ≥8-10 stable BG observations
- Intrawell dilemma
  - May have only 4-6 older, uncontaminated values per compliance well
  - Small sample sizes especially problematic for non-parametric tests
- Solution: periodically – but carefully – update BG data pool



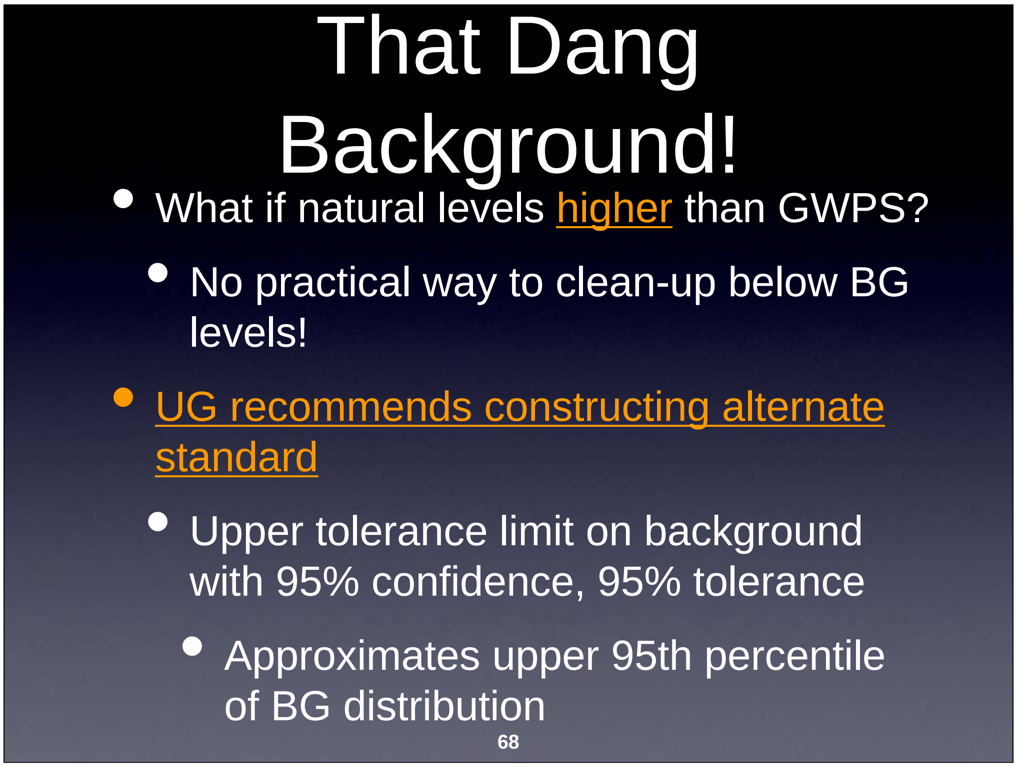
Updating Basics

- If no contamination is flagged
  - Every 2-3 years, check time series plot, run trend test
  - If no trend, compare newer data to current BG
  - Combine if comparable; recompute statistical limits (prediction, control)



Testing Compliance Standards

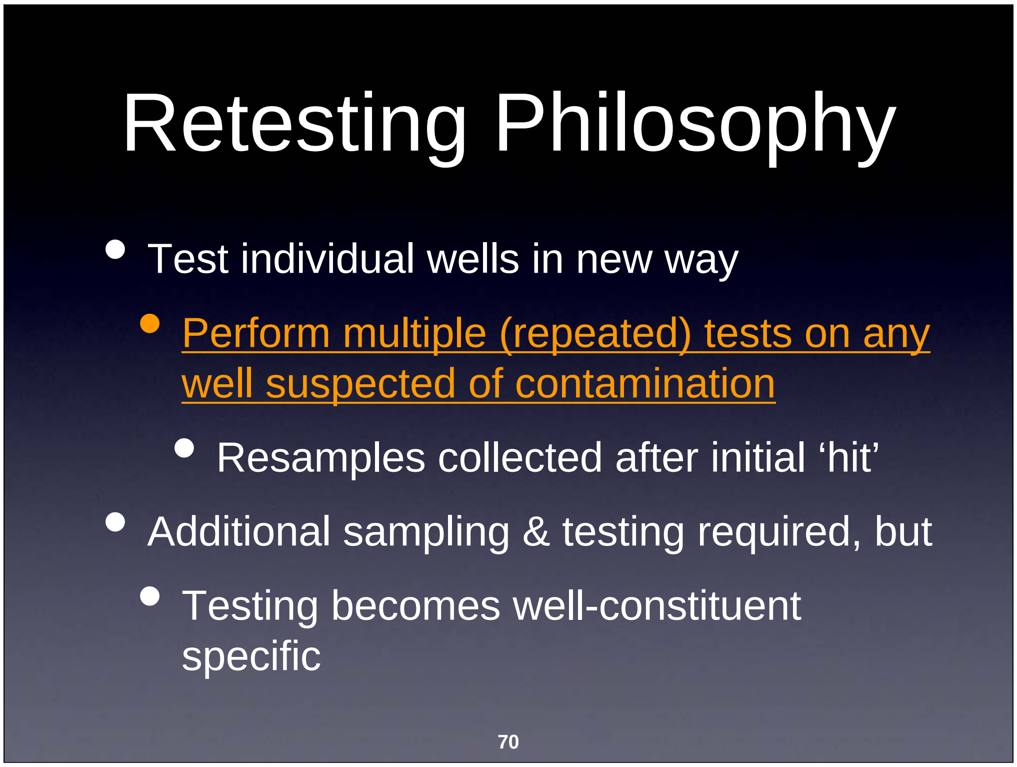
Schematic illustrates 3 possible outcomes for a confidence interval relative to a fixed standard; the left-hand case depicts a high level (at least 99%) of confidence that the population average is below the standard, while the right-hand case depicts the mean being above the standard, also with high confidence. Only the middle case is indeterminate with no clear statistically-based decision possible



That Dang Background!

- What if natural levels higher than GWPS?
  - No practical way to clean-up below BG levels!
- UG recommends constructing alternate standard
  - Upper tolerance limit on background with 95% confidence, 95% tolerance
    - Approximates upper 95th percentile of BG distribution

Retesting

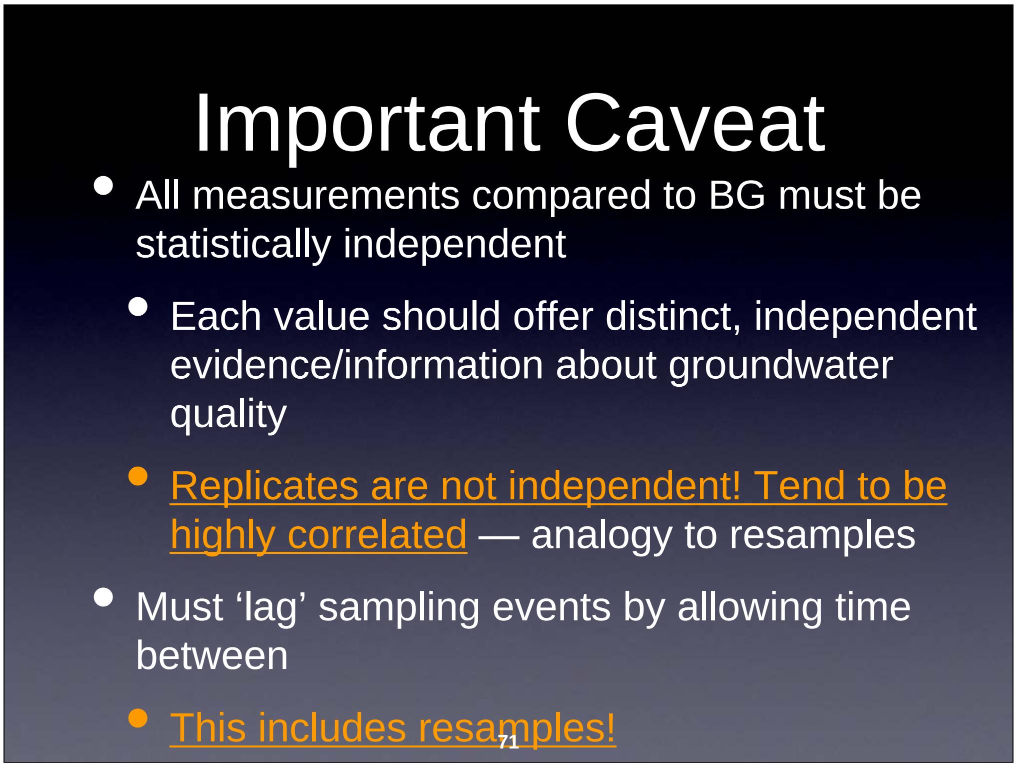


Retesting Philosophy

- Test individual wells in new way
  - Perform multiple (repeated) tests on any well suspected of contamination
    - Resamples collected after initial ‘hit’
- Additional sampling & testing required, but
  - Testing becomes well-constituent specific

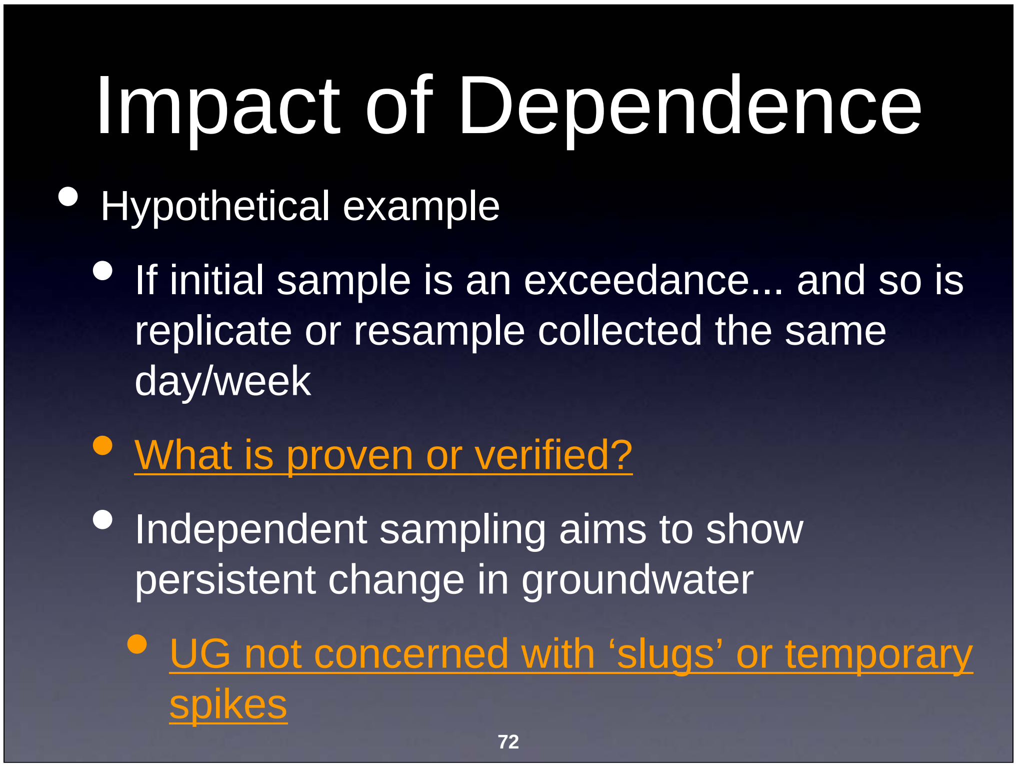
Retesting necessarily involves adding new data and information to the decision framework, but it does so in a highly efficient manner

Note that verification resampling has existed within RCRA regulations for at least the past 20 years; the difference then was that it was done on an ad-hoc basis as opposed to being part and parcel of the formal statistical framework



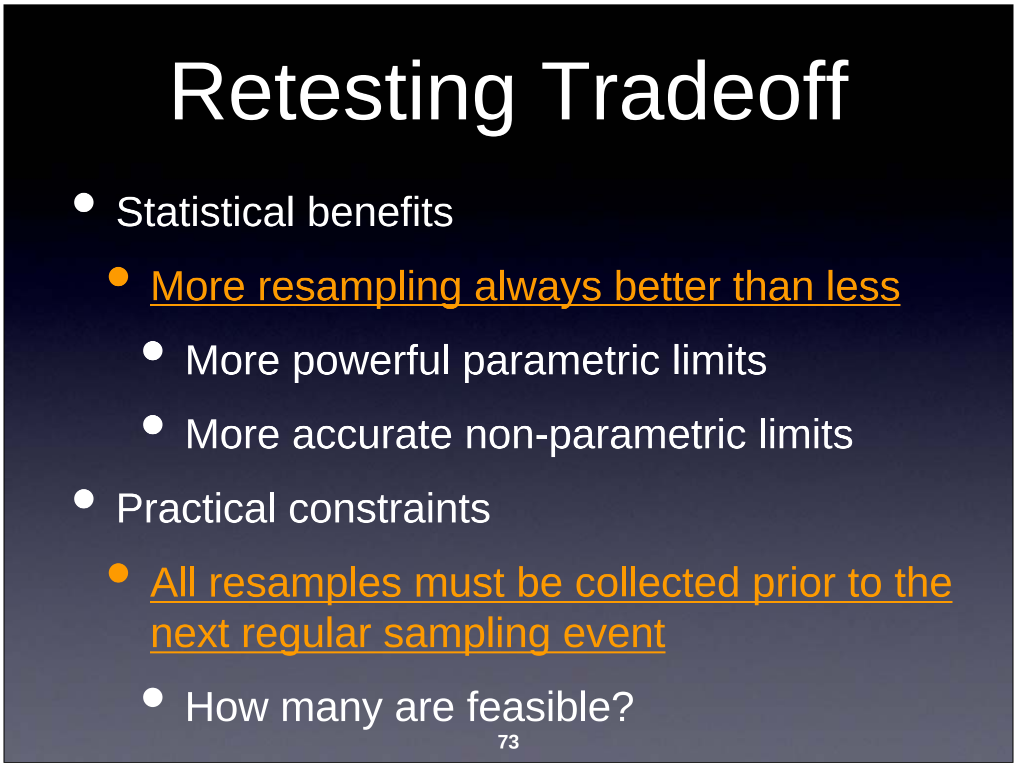
Important Caveat

- All measurements compared to BG must be statistically independent
  - Each value should offer distinct, independent evidence/information about groundwater quality
  - Replicates are not independent! Tend to be highly correlated — analogy to resamples
- Must ‘lag’ sampling events by allowing time between
  - This includes resamples!



Impact of Dependence

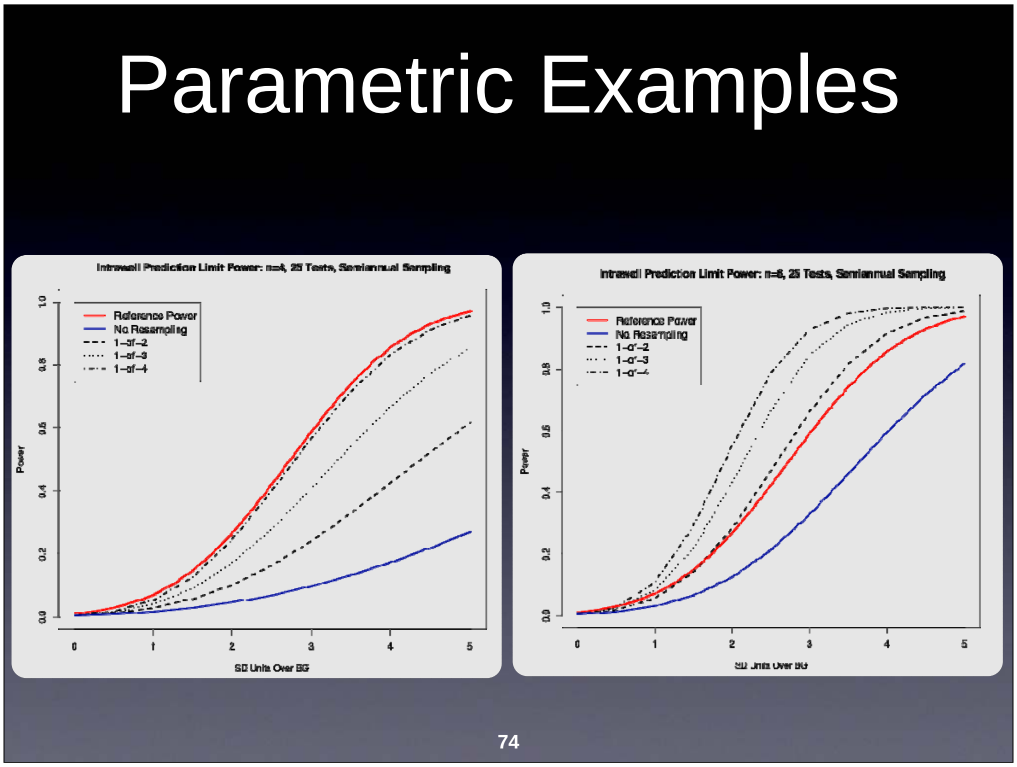
- Hypothetical example
  - If initial sample is an exceedance... and so is replicate or resample collected the same day/week
  - What is proven or verified?
  - Independent sampling aims to show persistent change in groundwater
    - UG not concerned with ‘slugs’ or temporary spikes



Retesting Tradeoff

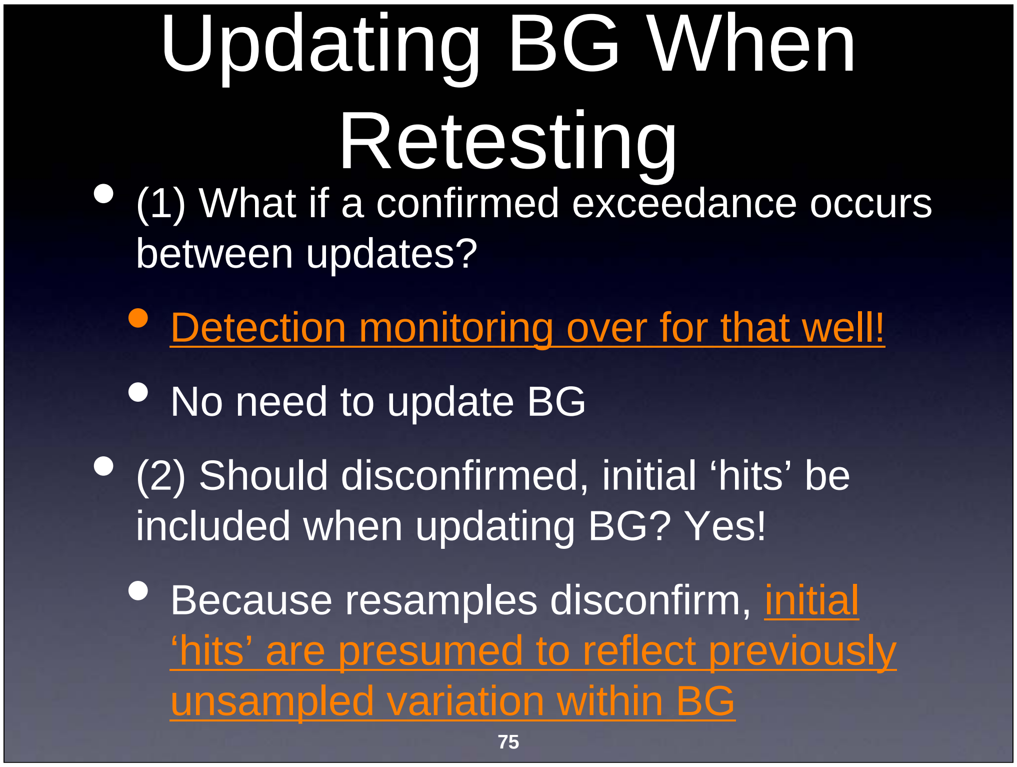
- Statistical benefits
  - More resampling always better than less
    - More powerful parametric limits
    - More accurate non-parametric limits
- Practical constraints
  - All resamples must be collected prior to the next regular sampling event
    - How many are feasible?

To avoid additional sampling cost, some have suggested that instead of collecting resamples between regular sampling events, that one simply use the next 1 to 3 regular sampling events as the resamples. The difficulties in this approach include the following: 1) practical and logistical — one would have to keep careful track as to which regular observations were being utilized as resamples and which were not, and at which wells; 2) statistical — in conjunction with the first reason and to avoid having to wait (for semi-annual sampling) perhaps a year or two to get a decision about a particular well, it has also been suggested that some regular sampling events might be able to serve ‘double duty’ as both the initial sample for that evaluation period AND one the resamples for a previous evaluation. The great difficulty here is the strong correlation that is induced between the regular samples and at least some of the resamples. Such correlation violates a key assumption of retesting that all the data being tested are statistically independent, and means that ‘all bets are off’ concerning advertised statistical power and false positive rates



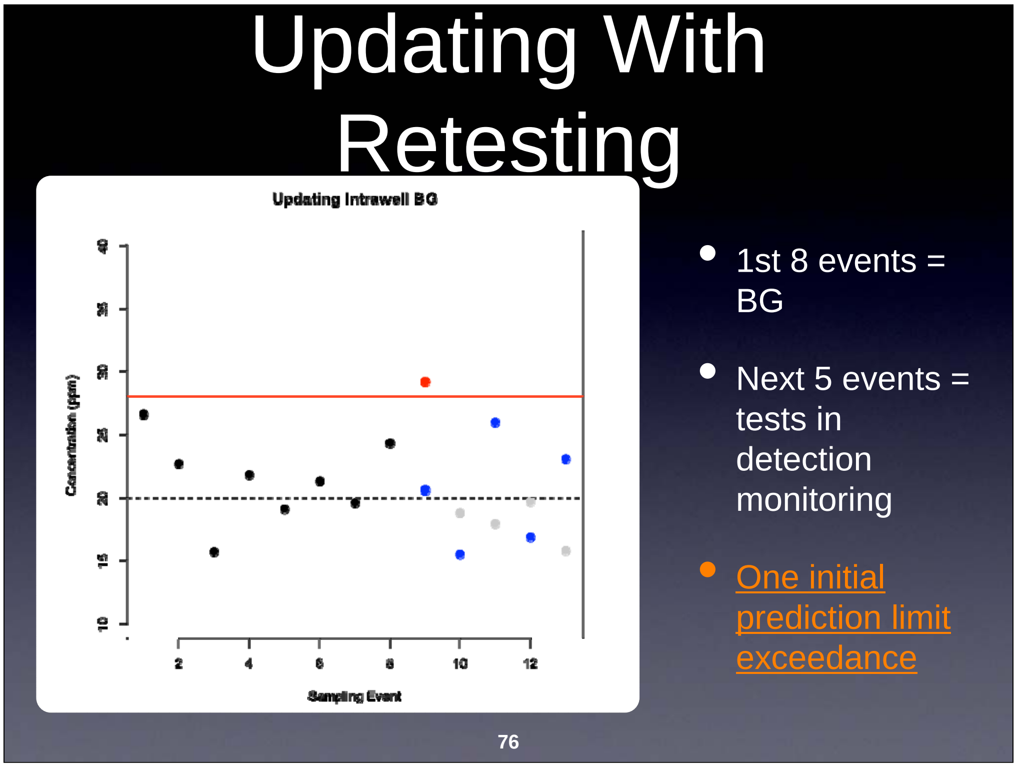
Parametric Examples

Each pane displays a comparison between not doing any resampling (blue) vs. using a 1-of-m plan; the EPA reference power curve is displayed in red on each graph; note that for these intrawell prediction limits at a site doing 25 tests per semi-annum, NONE of the 1-of-m plans is adequately powerful when there are only 4 intrawell background measurements per well (although the 1-of-4 plan comes quite close); by contrast, with n=8, even the 1-of-2 plan (that is, only 1 possible resample) is sufficiently powerful; note also that retesting substantially increases power in both panes compared to not doing any resampling



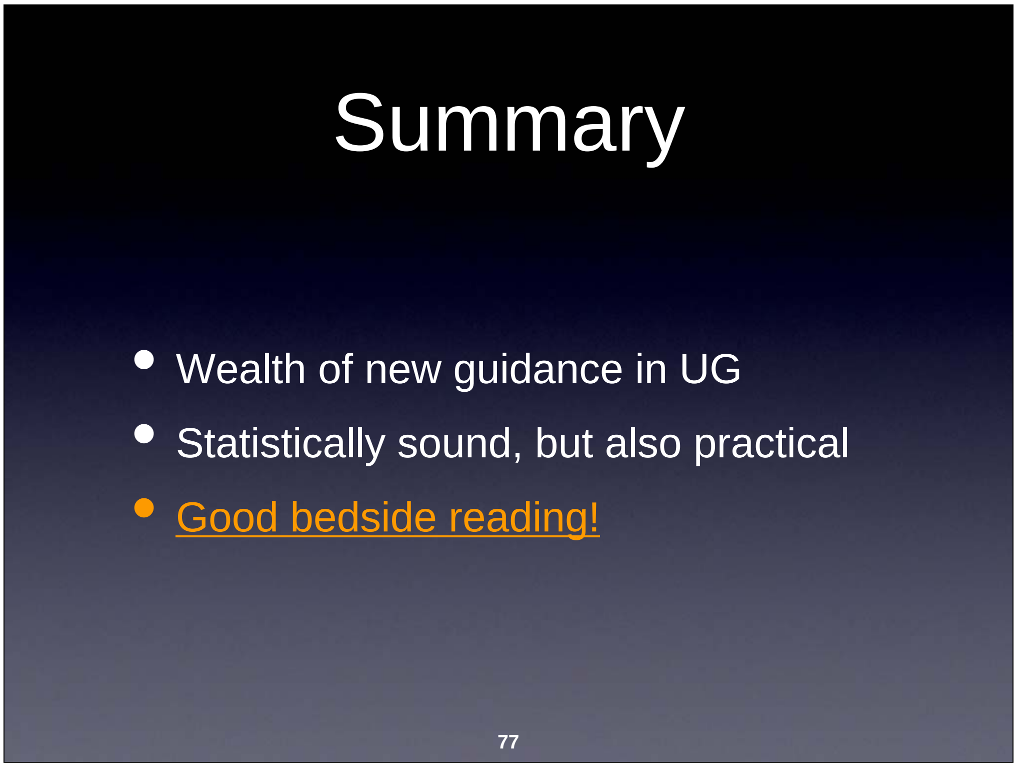
Updating BG When Retesting

- (1) What if a confirmed exceedance occurs between updates?
  - Detection monitoring over for that well!
  - No need to update BG
- (2) Should disconfirmed, initial ‘hits’ be included when updating BG? Yes!
  - Because resamples disconfirm, initial ‘hits’ are presumed to reflect previously unsampled variation within BG



Updating With Retesting

- 1st 8 events = BG
- Next 5 events = tests in detection monitoring
- One initial prediction limit exceedance



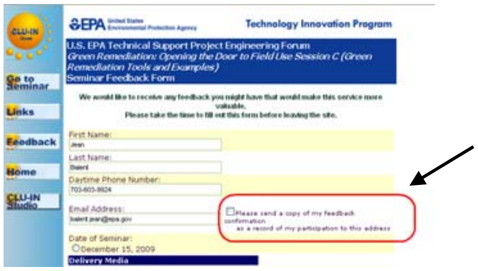
Summary

- Wealth of new guidance in UG
- Statistically sound, but also practical
- Good bedside reading!

77

Resources & Feedback

- To view a complete list of resources for this seminar, please visit the **Additional Resources**
- Please complete the **Feedback Form** to help ensure events like this are offered in the future


CLU-IN
Go to Seminar
Links
Feedback
Home
CLU-IN Studio
EPA United States Environmental Protection Agency
Technology Innovation Program

U.S. EPA Technical Support Project Engineering Forum
Green Remediation: Opening the Door to Field Use Session C (Green Remediation Tools and Examples)
Seminar Feedback Form

We would like to receive any feedback you might have that would make this service more valuable.
Please take the time to fill out this form before leaving the site.

First Name: Jean

Last Name: Balent

Daytime Phone Number: 703-603-9924

Email Address: balent.jean@epa.gov

□ Please send a copy of my feedback confirmation as a record of my participation to this address

Date of Seminar:
○ December 15, 2009

Delivery Media

Need confirmation of your participation today?

Fill out the feedback form and check box for confirmation email.