Review of "Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities: Unified Guidance"

Reviewed by Dennis R. Helsel, US Geological Survey. June, 2005

I fully accept attribution of this review to me, and welcome it being an open process.

Assigned Questions:

1. Does the Unified Guidance meet the stated objectives as a whole? In general,
a. Does the Unified Guidance effectively address the performance standards set forth in RCRA §264.97(i) and §258.53(h) and provide an effective framework for applying statistical methods for groundwater monitoring?

The Guidance is a definite improvement over what has gone before. However, it contains sections with substandard approaches.

b. Is the guidance presented in a manner that will be accessible to groundwater professionals with a limited background in statistics?
Please explain your answers and offer suggestions, as appropriate.

Some of the guidance is easy to implement with a limited statistics background. Other sections are not. There are complex discussions on power in Chapter 13 that should be eliminated. However the topic of power is very important, and an easier, more approachable section should be added. If the topic is ignored by State agencies due to the current discussion's complexity, which is likely, it will lead to substandard requirements for detection and compliance. Discussions on the non-central t distribution are not accessible to most readers, and should be moved to an appendix or deleted.

Most disturbing is the misuse of technical terms in the Guidance to seemingly impress those with a limited statistics background. Fabricating data is called "imputation". It is not imputation. Simulation of data with unrealistic characteristics is called a "model". Assuming all data below the detection limit to be at zero is called a "discrete probability model".

2. Overall, is the document well organized and cross-referenced in a manner that will help users apply the guidance? Please explain and offer suggestions, as appropriate.

Overall, the document is well organized. However, some reorganization will improve matters. Section 9.4 should be reorganized. Chapters 15 through 17 should be combined and reorganized, eliminating duplication of material.

Many more references need to be added to the guidance. Only occasional references point to what has gone before. This limits the ability of (especially non-statistical)

readers to find more detail elsewhere. It also violates good scientific practice, leaving the impression that original work is being presented here. No references are given when discussing the effects of transformations. No references are given when comparing nonparametric and parametric tests. No references are given for any of the formula presented. Especially lacking are references to environmental statistics texts and industry-standard introductory texts on statistics. If published as a commercial textbook, this guidance would be considered as containing plagiarism.

Also disturbing is the lack of references in the statistical tables of the Appendix. If these result entirely from new, original generation of numbers by the authors, then fine. If however the tabled numbers come from other textbooks and papers, those sources MUST be referenced here. And note that textbook publishers do not allow tables to be copied from their books with just a reference. As opposed to articles, textbook publishers charge a fee to use information from their published tables. If this is not sorted out now, later a publisher may cause some trouble for you. Of course, if all of these come from sources on the web, that's fair game as long as the URL for the source is referenced.

3. For each of the five major sections of the Unified Guidance, please address the following questions :
a. Does each section of the guidance meet the stated objectives described in the Introduction to the Charge, above? Please explain.

Sections I to III generally meet the stated objective. Sections IV and V do not. Section IV contains recommendations at odds with the previous three. It defines 'trends' differently, measures trend differently, and uses different methods for nondetect data. Its three chapters are repetitive, and so not well organized, and could be greatly condensed, discussing a topic once instead of three separate times. Section V contains 'research' results that have not been presented elsewhere, and so have not received review other than in this process. Review of the technical content of this material should be performed by persons with the technical expertise found in journal reviews. I have given review comments in the detailed comments section, below. It appears to me that the research results presented in Section V are technically flawed, and so the recommendations based on this material is likely to be wrong.

b. Does each section cover an appropriate range of topics? Are there any key topics that are missing or that should be emphasized or described in further detail? Please explain.

The ANOVA section excludes discussion of the Kruskal-Wallis test as a nonparametric option. The reasoning behind this is not given. It would seem that the same approach that led to recommending use of the Wilcoxon rank-sum test for two-group comparisons would also apply to the many-group ANOVA context, whether applied to spatial or temporal differences. Either the K-W procedure should be added, or a clear explanation of why not should be added. The current recommendation to use transformations followed by ANOVA is flawed, as there is often not a single transformation that makes all groups look like a normal distribution. A transformation often makes one group more normal, while another less so.

Tests for adherence to a distribution can be performed for censored data. Those methods are missing here. The current description of probability plots for censored data is wrong, and plots will be incorrect. The statement that distributional testing with censored data cannot be done is also wrong, and should be changed.

c. Is the material in each section organized and presented in a clear and concise manner? Please explain.

This material is definitely NOT concise. Concepts are repeated many times over. Whole paragraphs in one section repeat the same thoughts that occur in previous sections. The tone is often one of a conversation, sometimes rambling, more than that of a handbook. The document could be condensed by about 20 to 25 % by a strong editor.

4. Are the methods, approaches, and strategies described in the Unified Guidance technically valid and accurately interpreted, described, and applied in a groundwater monitoring context?
Please comment on specific methods, approaches, and strategies, as appropriate.

The approaches and strategies for nondetect data are definitely invalid, and will lead to incorrect decisions. They are based on untested alterations ('fixups') to 1950s methods. They are "seat of the pants". Industry-standard methods (Kaplan-Meier methods) exist that can be presented at a level suitable for this document. But those better methods are not found here.

"Insider censoring" of low-level data is advocated in Chapter 10. This method will produce biased results. This occurs when data measured as below the MDL are reported as <QL, the quantitation limit. This recommendation should be scrapped.

Land's method for computing an upper confidence bound is still recommended. It should be dropped. Several EPA publications, including ones cited in the Guidance, state that other methods should be used instead. Land's method is especially poor for small data sets, and for data where the true distribution is unknown. This is precisely the type of data that it will be applied to if the Guidance recommendations are followed.

The Guidance improves greatly on past information and guidance in the RCRA program. Of special note as improvements are prediction limits, one-sided tests, testing for normality, and discussions of transformation bias and its effects.

5. In your opinion, what are the weakest and strongest aspects of the various sections, chapters and/or recommended methods? Please make suggestions on how the weakest parts can be strengthened.

One of the weakest areas is on power. There is much space devoted to it, and a number of (sometimes repeated) equations. But it is too complex to be helpful to a non-statistician, who is probably the one making decisions on requirements. The result will

likely be that insufficient samples will be taken to detect contamination in a significant percentage of cases. Since 'no contamination' is assumed until proven otherwise, 8 to 10 samples is rarely sufficient to definitively prove that contamination has occurred. The discussions on power should include worked-out examples of typical cases. A table that lists power resulting from a series of typical sample sizes should be added for each type of test, to illustrate how many samples are likely to be required. This means, of course, that estimates of delta (expected contamination levels) must be gotten from somewhere, but this should be available from existing studies and experience in the program by now. In particular, delta in a percentage context, a 10 or 20% increase over background should be detected, is on a scale useful to the regulatory agency. So a table that states "to detect a 10% increase, you need 30 samples for 80% power. For a 20% increase, you need 25 samples for 80% power", etc.

The second weak area is the discussion of nondetects. Aitchison's method as implemented here is not Aitchison's method, but a smokescreen for substitution of one-half the detection limit. Numerous studies going back to the eighties have shown that this substitution does not work well. Cohen's method is noted to work only for one detection limit. In that situation, it is an acceptable alternative if tables are better than using software. It should be stated clearly, however, that the method fails badly when more than one detection limit is present. The simulations presented in the Appendix are flawed, and should not be used to bolster these two procedures. Any time that major recommendations are made, supporting work such as these simulations should first be colleague reviewed. These simulations should have been submitted for publication in a qualified journal, reviewed and accepted before using them to bolster the recommendations. My review of the simulations, below, can be summarized to say that they are flawed, and were based on unrealistic simulated data, data that does not look like what is found 'in the ground'. The simulated data mirror the assumptions of the tests used. Alternative tests found to be better by others were not considered. So recommendations to use the Cohen and Aitchison methods appear to be supported much more than they actually are. The simulation means nothing for how well these two tests might work with actual concentration data that look unlike the data generated here.

A less severe weakness is that for several of the worked examples, data sets were chosen that are too simplistic. Obvious issues like not being either normal or lognormal, or having multiple detection limits, are not included. So a user may find that they understand the computations, but cannot perform the computations on their data because the guidance does not tell them how to.

6. Are you aware of any other significant methods, approaches or strategies that are relevant and should be included in the document? Please explain, offer suggestions regarding where and how the methods/approaches/strategies could be incorporated, and provide relevant citations.

Kaplan-Meier methods are the standard procedures for handling censored data in medical and industrial statistics. Like other nonparametric methods, they are a counting procedure and are not hard to do by hand. They are certainly easier to explain than some

of the current topics like power calculations or non-central t presented in the guidance. There should be no impediment for using them here instead of the Aitchison and Cohen methods. An important reference on the topic is Helsel (2005), Nondetects and Data Analysis, Wiley. A readable reference from the medical statistics community is Klein and Moeschberger (2003), Survival Analysis, Springer.

As mentioned elsewhere, Kruskal-Wallis and censored probability plots (done correctly and for multiple detection limits) should be added.

Specific Topics
1. The Unified Guidance presents a comprehensive approach to address the multiple comparisons problem in detection monitoring. Both sitewide cumulative false positive and false negative (power) errors are addressed, primarily through the use of prediction limit retesting strategies. Is this approach reasonable and sound? Please explain and offer any suggestions, as appropriate.

The current approach appears reasonable and sound.

2. The Unified Guidance concludes that similar cumulative false positive errors cannot be realistically defined for compliance/corrective action testing against a fixed standard. Two major recommendations are provided: 1) a priori power criteria to allow for consistent ability to detect increases above a standard under conditions of fixed or small sample sizes, and 2) aggregation of annual data to enhance both power and single-test false positive errors. For corrective action testing, enhancing power is left to the discretion of the facility beyond aggregating annual data, and a predetermined single-test false positive is recommended. Is this approach reasonable and sound? Please explain and offer any suggestions as appropriate.

This topic is out of my area of expertise.

3. Please identify any other recommendations that represent a revision and/or enhancement to current guidance and practice and that have the potential to significantly affect groundwater monitoring under RCRA or other environmental programs. For each topic identified, please answer the following questions:
a. Are the recommendations appropriate and reasonable given available methods, documented experience, and current practice? Please explain
b. Does the document provide adequate guidance to help owners and operators, Regional and State regulators, and others put these recommendations into practice? Please explain and offer suggestions, as appropriate.
4. Are the statistical method summaries and flowcharts in Chapter 5 useful, and do they provide clear guidance for potential users?
5. Is the software program for Chapter 13 non-parametric prediction limit testing useful and accurate?

(answers to these questions are included in the specific comments, below)

Specific comments on the text

Page	Comment
3-14	The term “needle in the haystack” isn’t applicable here. A 3 sigma increase is hardly a “needle”. This may be unnecessarily adversarial. Delete the term here and when its used later.
3-26	Figure 3-4 is incorrectly labeled as Figure 3-5.
4-2	How about using “convicting the innocent” instead of “hanging’?
	It is true that the null hypothesis is favored, but it is not true that the null hypothesis is always defined as ‘in compliance until shown otherwise”. Several regulatory programs assume non-compliance until shown otherwise, as you do for compliance monitoring. This provides a better incentive for collection of sufficient sample sizes, and adequate power, than does this section of the Guidance. It is easily done by reversing the inequalities from your formulation. I am not suggesting you make that radical of a change at this point in the program, but it would be a good addition to specifically state that this alternative setup is possible, and EPA has chosen to do what you have outlined. It should include a recommendation for regulators to take power considerations very seriously.
4-3	the phrase “under the null hypothesis model” should be explained, or changed to “when the null hypothesis is true”. “Under the null hypothesis” is used throughout the Guidance. If this document is actually targeted to persons who may not have extensive statistical training, then the “under” phrase will be jargon and not understood. Either change it to something like the suggestion above, or at least explain it at its first usage.
4-11	Place the term “Power” on the upper left cell of Figure 4-4.
5-8	More detail is given in comments on chapter 10, but “simple substitution” is NOT imputation. Imputation methods use some sort of model for the data to derive values, and result in values that are not all identical with one another. There is no model in your substitution -- it is like calling 42 the answer to the Question of Life (Hitchhikers Guide to the Galaxy).
	Why are there no nonparametric analogues to ANOVA (Kruskal-Wallis) listed in the Chapter 9 methods? This would give a test for spatial variability across groups without assuming normality or equal variance. Why include nonparametric methods in Chapter 11, but not in 9?
5-18	The standard form of Levene’s test subtracts off the median rather than the mean prior to performing ANOVA. Subtracting the mean was an option found to be less robust by the developers. Provide a justification, or at

least a reference, for using this form if you decide to continue using it. Minitab's implementation of Levene's test subtracts off the median, for example.

5-29 Cohen's method can only be used when there is one detection limit. This should be listed in either U**nderlying assumptions** or **When to use**. It can be used when data are re-censored to the highest of multiple detection limits, but this introduces even more error, and is unnecessary, as there are better alternatives.

5-30 Your version of Aitchison's method states that all nondetects have the same value. This is not a "discrete distribution". It is a plug-in value, and very difficult to justify. No chemicals that I am aware of, when detection limits have been lowered, have produced all the newly detectable data sitting all at the same number. Aitchison's method was developed for economic data where zeros were plausible. It was shown to be a poor estimator of the mean and other summary stats for environmental data by Gilliom and Helsel in the 1980s. USEPA modified it in guidance around 1990, calling it the "modified Aitchison method" by plugging in values at the detection limits instead of zero. From published simulations by Hinton (1993), that was shown to be a poor method. Now you are recommending yet another plug-in at one-half DL, re-using the same name. If you continue to recommend this procedure, at least rename it to distinguish it from the previously discredited modified Aitchison's method. In fact, it is so different from Aitchison's original method (which used MLE to estimate stats for the detect portion) that you should simply call it what it is, the "substitute one-half dl method". It is not Aitchison's method. It is modified only in the sense that the original method was dropped and something else done instead. It is a substitution method using whatever fraction of the detection limit is in vogue this time around.

5-33 The Wilcoxon test works well for data with one detection limit. Make this explicit. It can be used on data with multiple detection limits by re-censoring all values below the highest detection limit as tied. This loses some information, but perhaps not a lot depending on the pattern of detection limits and data. It avoids the biggest problem with your recommended substitution methods, that is, adding a signal to the data that was not there to begin with.

5-42 Another substitution, here of one-half the detection limit for Poisson limits. The air quality and health people use 1 over the square root of 2, which is about 0.7. Simulations in the nineties showed that if data were from a lognormal distribution, using 0.7 DL was the best single plug-in to duplicate the mean. Are you sure you want to use a plug-in value? This opens you up to complaints (and lawsuits) that you used the wrong one.

None of them work well in comparison to better methods. More on substitution later in comments on chapter 10.

5-45 It is not made clear that the Mann-Kendall test is just Kendall's tau correlation coefficient, and thus duplicate in function to Spearman's coefficient. Kendall's tau normal approximation is good for n>10, while Spearman's requires 20 or more observations. You appear to switch those two qualifications in the Tables referred to. Why? What reference backs that up? More on this later.

5-46 You state that Sen's slope needs no special adjustment for ties, implying that nondetects are easily handled. How? When computing the slope of delta Y over delta X, and Y is a 10 versus a <5, what is the slope? Methods have been proposed for this situation, but you do not cite or describe them. You gloss over this far too quickly.

5-49 "level of statistical variation". Is there any other kind? Drop the word statistical.

You do have a number of caveats about Land's method, but why recommend it at all? Other USEPA documents for other programs have shown it to produce unreasonably high values for the upper confidence limit. It works very poorly for small to medium data sets, so large amounts of data are needed to justify its use. See papers by Anita Singh, including some USEPA treatises. I recommend dropping this as a recommended method.

6-1 The issue for water quality data and a normal distribution, even background data, is the possibility that for a normal distribution, values can go negative. This makes a normal a poor guess for low-level data. Since with small sample sizes it is difficult to reject the assumed distribution, assumption of a normal is generally not good practice. Your later emphasis on testing first is a much better approach.

6-2 There is no such thing as a "non-parametric distribution". Change to something like "use of nonparametric methods". This phrase is also used in the Appendix materials. Change it.

"tests on the mean are generally more robust…" It is extremely important to define your terms here. The robustness referred to is the formal statistical definition, the avoidance of Type I errors. This is not the issue. The issue is that normal-theory tests have very low power when applied to non-normal data. So Type II error rates are huge. The non-statistical reader will assume that 'robust' means the tests don't generally make mistakes, and are generally applicable. "Robust" coffee is strong and full-bodied, so these tests must be strong and potent as well. This non-

technical definition of robustness is not true. The result will be large Type II error rates. Contamination will go undetected that should not.

6-6 to 6-11 After a long discussion, the conclusion that data should first be tested using one of the recommended methods in order to determine whether a normal or lognormal distribution is more likely, makes much sense. This is much better than a blind assumption of either a normal or lognormal distribution. The statement to do the testing should be made at the outset of this discussion, before the simulation results are discussed. Than shorten all of the qualifications that obscure this important guidance.

When there are censored values, use of a censored probability plot (and associated test, which is nowhere described in the UG but certainly should be) should be strongly recommended at the outset. Use of the tests up front make much of the argument of whether to assume one or other distribution moot. That is a very good thing, as the arguments presented here for assuming a normal are based on faulty assumptions. For censored data as for uncensored data, do the plots first and don't assume one or the other distribution blindly.

The simulations are unrealistic. One issue is acknowledged within the discussion – a normal distribution with a CV greater than 0.3 will produce a significant proportion of negative values. Use of such a distribution in simulations produces results which won't correspond to what happens in reality. It is pointless to do so. Compounding this is the 'imputation' (misuse of the term) or plugging in of a single small positive value for generated negative values, as if they were nondetects. This produces a distribution that is not a normal distribution. So stating that this condition represents a normal, and using it for justifying the assumption of a normal distribution, is absurd. The simulation results for the normal distribution presented above a CV of 0.3 are invalid, and should not be presented nor used for decisions as to which distribution has more penalty when assumed.

Your recommendations on page 6-10, "In summary…..", don't make sense based on your simulation results. It is not clear how you came to them, based on the previous results. Based on your own discussion, your recommendations should be dramatically changed. The recommendations, based on your findings, should be
1) for n>10, don't assume a distribution. Test to see whether a normal or lognormal fits the best.
2) for n <=10, if there is to be no retest, assume a lognormal. A much better practice would be to assume a normal but perform a retest.

Section 6 You do have a few citations in this section, but not many. In general, this Guidance is woefully missing references to the good work of others that is

drawn upon here. What about general environmental stat texts like Millard's "Environmental Statistics with S-Plus"?.

7-1 Boxplots are not a method for testing homogeneity of variance. You should tighten up your description here. Later in section 9 you state that you are using methods to explore and test for….This is far better. Boxplots do the explore part.

Section 7 It would be helpful to list which software packages compute each method recommended. So for example, which packages compute Levene's test? A useful table in the appendix would list the methods recommended here, and have a check mark for the commercial software that does them. Also check your definition – as stated earlier, the standard methods I have seen for Levene's test subtracts off the median rather than the mean.

8-3,4 Your discussion of dealing with nondetects gives the reader no concrete methods for dealing with multiple detection limits. Your acceptance of "nominally assigned values" (perhaps "nominally intelligent methods"?) leaves open the option of real error. Data should not include substituted values. Instead, censored probability plots should be used. Commercial software has no problem drawing censored Prob-plots for multiple detection limits. No substitution is needed, or wanted. Point people to these methods rather than some "quick and dirty" fix up. Most of the appropriate methods for censored data come from the drug testing industry. Would you want a family member to be taking a prescription whose safety test included 20% or more of numbers that were simply made up? I wouldn't! Why should environmental work be any different?

When a censored probability plot is to be used after labeling outliers as "greater-thans", what value is the outlier greater than? Give more specific guidance on how to do this.

Section 9.3.2 In addition to ANOVA, why isn't the Kruskal-Wallis test listed as an alternative when data are not normally distributed? You recommend taking logs and running ANOVA on the logs. That tests for differences in mean logarithms. If the logs were normally distributed, as one would hope, this is also a test for differences in the median logarithms. Translating back to original concentration units, the test on logs is a test for whether the geometric means differ from a ratio of 1. If the logs are normally distributed, it is also a test for whether the median concentrations differ from a ratio of 1. (I later read your discussion of the same thing, but its not evident here). Parametric tests using any transformation that produces a normal distribution, logs or otherwise, is a test for differences in group medians in the original units (and not group means). The K-W test is just a clearer way to test for differences in medians. The impression here is that you use ANOVA because it has the word "mean" within it,

even though after transforming data the means are no longer being tested. You should cite the Kruskal-Wallis test as a valid alternative, and a time saver, to test for group differences without searching for the best transformation.

Section 9.4.2 Why use this test when the adjustment for ties is missing, given that you expect many ties to occur? The alternatives listed do not include the simple nonparametric correlation coefficients to be introduced later, Spearman's and Kendall's. Either can be used to test for serial correlation between the lag-1 pairs, and both handle ties with an appropriate correction. My recommendation is that for simplicity as well as an available tie correction, you switch to using one of the nonparametric correlation coefficients instead of this test.

Section 9.4.3 Again, add the Kruskal-Wallis test as appropriate when the data are not normally distributed. It is far less work than searching for a suitable transformation. With multiple groups, a transformation that normalizes one group may make another appear even less normal. So a nonparametric test often has greater power due to the difficulty of finding a suitable transformation. Perhaps the reason ANOVA is emphasized is to get an estimate of the pooled standard deviation? If so, make this objective clear in the text. You aren't necessarily trying to test for differences as much as estimate a pooled statistic.

Section 9.4.4 You've jumped from testing for serial correlation (9.4.2) to testing for temporal differences (9.4.3), and back to adjusting for serial correlation. The two different objectives of sections 9.4.2 and 9.4.3 will easily get confused. Move this section to become 9.4.3, so first test for it, then adjust for it in response, before moving on to a different objective. Make the shift in objectives more clear to the reader once you start the new topic of looking for temporal differences. So the ANOVA section for temporal differences will become 9.4.4, and goes with the adjustments of section 9.4.5. Separate and be consistent in objectives.

10-1 Before basing guidance on simulation studies, it is first necessary to get those studies peer-reviewed. This would be best done in a journal rather than within the review process for the guidance document itself, which may be done by generalists rather than persons best qualified to review the technical detail of the simulations.

10-2 Your recommendation of reporting the nondetects as less than the QL, and then using the estimated values that are below the QL as qualified, produces a positive bias in the data. All subsequent analyses, including computing means and confidence intervals, will be based on positively biased data. This is called "insider censoring". Two solutions are either to use the MDL as the censoring level (what the lab actually reports), or

censor all observations measured below the QL as <QL and not use the estimated values. Change your recommendation to one of these alternatives. Your current recommendations are incorrect.

10-4 Terms like "a reasonable proportion of the time" lead to controversy. If EPA has an estimate of what is reasonable, it should be included in the guidance. If not, expect controversy.

10-8 Your simulation results appear 'reasonable'. Your conclusion that "none of these tests is likely to offer much chance…" is really not a function of the test. It is that these small sample sizes are not likely to offer much chance of finding a signal, if present. Rewrite this conclusion. As it now reads, someone might hope for a 'better test', when in fact the issue is lack of power due to small samples. Rewrite to say that these simulations prove that sample sizes this small are insufficient for a monitoring or detection process.

10-10 Based on additional experience and research, which is published in the open literature, I disagree with your statement that up to 20% of data can be substituted with made up numbers and it has no effect. One item rarely done in simulations, if this is the basis for your statement, is that detection limits vary with interferences, different labs, etc. So the simulations should be based on multiple detection limits where the values of the detection limits are somewhat randomly determined, and occasionally quite high, rather than using a consistent detection limit at a low, 'research' level. If these more realistic scenarios are used, a false signal is easily 'imputed' into the data that was not previously there. This signal depends on the pattern of detection limits, not on what was in the samples. It interferes with obtaining the correct results for hypothesis tests.

10-11 The Kruskal-Wallis test is simply an extension of the rank-sum test to more than 2 groups. You recommend the rank-sum test as the most appropriate test to use in the two-group case, yet reject use of the Kruskal-Wallis test for more than 2 groups. Your reasoning, even with the text on this page, escapes me. Your first argument is that "identifying temporal correlations is likely to be extremely difficult if a rank-based test must be employed….". Why this is so is not stated. Rank based nonparametric correlation coefficients are used all the time to test for serial correlation. Your second argument is that a KW test might be sufficiently powerful to see differences, even with large % of nondetects, though the magnitude of differences might not prevent pooling. So the solution is to use a less powerful test? This makes no sense to me. Why not specify a percent difference in spatial variation, below which pooling is acceptable?

Your recommend parametric tests when nondetects are "concentrated heavily" in one or more groups, after stating that substitution (the only

way you've identified for running parametric tests on censored data) should not be done with greater than 20% nondetects. This is certainly contradictory. You will get a variety of inconsistent, seat of the pants approaches as a result of recommendations of this type. Instead, use one of the better parametric and nonparametric approaches out there for handling nondetects.

10-12 You are again recommending substitution of one-half DL ("one could" do this), which is NOT a 'simple imputation' strategy, but a fabrication strategy. Then you acknowledge that others have found that this method is inadequate, and produces poor results. So stop recommending it, or even mentioning it as a possibility, otherwise people will continue doing it and claiming that the Guidance allows it, or recommends it, or whatever. If you want them to use Cohen's method, point them to it without allowing something worse. Aitchison's as you've defined it is just substitution, and is already worse.

10-14 Aitchison's (original) method was evaluated by Gilliom and Helsel (1986) and found to perform quite poorly for estimating both the mean and standard deviation. It is in essence a substitution of zeros. Why are you still recommending it? Other procedures work far better. The "modified Aitchison's method" of substituting the detection limit was evaluated by Hinton (1993) and found to work poorly. Why do you think your modification, fabricating with a different number, will fare better than either of these other two?

Cohen's method was designed for use with one detection limit only. If more than one limit exists within a data set, all data below the highest limit must be censored as less-than that limit, in order to use the method. This is a great loss in information. You have just rejected using the test of proportions because of a similar loss of information, but now you adopt it here. Why are you still recommending Cohen's method?

10-15 So here you explicitly recommend substitution of one-half multiple DLs. This introduces a signal that is not in the data. You are making up data, then analyzing it. This is not imputation, but fabrication. It should never be recommended by EPA. There are good alternative methods.

You don't explicitly state it here, but Cohen's method will not handle data with 3, or 5, different detection limits. It will not recognize the difference between data containing <1s, <3s, and <5s versus another data set that contains <1s, <5s and <10s. In short, for the common occurrence of multiple detection limits, Cohen's method is not applicable.

Did McNichols and Davis study the validity of t-tests using Cohen's method when there are multiple limits? Do you feel capable of extending

their results to that case without any grounds for doing so? Did they account for data reported as <10 (the QL) when it was in fact measured as <5 (the MDL), as you are proposing to do? When all data measured as less than 5 receive a substituted value of one-half QL, or 5? Did they include this bias in their simulations? I think not. You are using their simulations to support a method that is far worse than what was supportable.

10-17 Your censored probability plots are overly simplistic, and not the way that commercial software would do them. In your method, all <10s (as an example) would receive plotting positions lower than a detected 3 or 5 or 8. Yet in reality they may not be lower. The standard methods for constructing censored probability plots account for some probability that a <10 could be higher than a 5. The standard methods use either Kaplan-Meier, maximum likelihood, or Helsel-Cohn methods, the latter being the most common so far in environmental studies. Your recommendation should be to use one of these three methods.

What if censored, lognormal data were plotted on a censored normal probability plot and the resulting curve seen? Your recommendation would be to use Aitchison's method, because of the curvature ("significant bends and curves"). Yet its origin is due to a different overall distribution, rather than points to a different way to model the nondetects. Have you simulated the errors resulting from choosing Aitchison's method on the wrong distribution as a result of your recommendations? Instead of doing so, use standard censored probability plots and do not use either Cohen's or your modified Aitchison's methods.

10-18 So what happens if there were a 4 measured during the time the detection limit was 3? Is it assigned a value higher than all the <5s? You've chosen a convenient data set, but so unrealistic as to not be helpful to the reader.

10-24 You have cited Cohen's method on many pages. This is the first acknowledgement that it only works for one detection limit. Imagine the disappointment the reader will have when they get to here and realize this method that you propose in detail, with examples worked out, won't apply to their data. This is the norm, not an exception! Most of your readers will be facing multiple limits. They get one paragraph that gives references on how others have dealt with this, but no worked out examples, no discussion of how those methods compare with Cohen's method, and the recommendation to consult with a professional statistician! Better methods than Cohen's exist, they are available in commercial and free software (software for the Helsel-Cohn method has been freely available online for years), and you could put an example of their use here without much difficulty.

10-25 You are flying by the seat of your pants here. How do you know that the median of detection limits can be used? For your data, certainly the higher limit could, as all <3s are also <5. There is no justification for using a 4. And if there were detected 3s or 4s?

10-26 Substituting one-half the DL may be "simple substitution", but it is not "imputation". Imputation implies there is some theory or model behind the estimation of values. There is none here, as shown by arbitrary modifications of the original substitution of zeros to substituting the DL, to now using one-half DL. It is all fabrication of data.

Your simulations should have been colleague-reviewed before using them to justify recommendations. Therefore the validity of your recommendations cannot be verified. Recommendations with this much weight, which will influence the spending of millions of dollars, should not be based on unverified studies. Did your simulations account for your recommendation to report <MDLs as <QLs, for example? So that values substituted are the maximum of the value they could have been measured at on the instrument (if measured as higher, they would be J values and their numbers used)? Did you include in the simulations the real probability that the distribution might be mis-specified? Did you include the possibility that seemingly randomly, interferences or labs might have upped detection limits, so that a simulated 0.5 in some instances might have been called a <1, and in others a <8, resulting in very different substituted values? I reviewed your simulations under comments for the Appendix. I don't believe they support your recommendations here. They are flawed.

10-27 "works best"? Better than which alternatives? Under which scenarios?

Section 10.5 Aitchison's method computed the mean and standard deviation of the detected observations by maximum likelihood rather than using sample statistics. So you are modifying his method even more so than indicated. Acknowledge this additional modification in your description. Since you have changed both the value assumed for the nondetect spike, and the way to compute stats for the detects, you really don't have anything of his method left. Once you take a Lexus and swap out its chassis, engine, transmission and body, substituting one from a Ford, you no longer have a Lexus. Your 'Aitchison's method' is really nothing other than substitution. You should just call it substitution! You have modified Aitchison's method so much that you no longer have his method. While "modified Aitchison" sounds more responsible than "substitution of one-half QL", the latter is a much more accurate name.

You don't state what the lower confidence bound on beta is based on. What is Table 10-2 of the Appendix? Is this a binomial table? What

reference is there for computing a lower bound on this proportion? Why was a 50% interval chosen here? What are the consequences of a 50% probability that the proportion is lower than the lower bound? Did you take that probability into account in your simulations on this method?

Section 11.3 State whether the Welch's t-test is performed by standard statistical software. Which do this, and are there other varieties of t-tests that an unsuspecting user could end up doing? In particular, state that users should not check the option to assume equal variances when running t-test software. This would not be the Welch's form of the test.

Section 11.3.5 It is unclear why this is here. It is much more complex than the rest of the guidance. Indeed, the guidance has avoided anything much more complex than computing a mean to this point. Now suddenly the non-central t distribution is discussed. It is unlikely that this section will be used by the target audience. This section can be deleted.

11-16 The rank-sum test 'loses' information only when that information is contained in the data. It is contained in the data only when the parametric assumptions are satisfied. So when data are non-normal, the rank-sum test is 'losing' only misinformation, not information. Change your discussion to reflect this. Lehmann's book should make some reference to this.

11-17 Your preference for the t-test (with substitutions) when nondetects are fewer than 20% is curious. What advantage can there be in using a poor method over a method which makes efficient use of the data and makes no arbitrary substitutions, even for that 20%? Do you have simulations that show when one-half QL is substituted for values measured as below the MDL, and where multiple QLs add in an arbitrary signal, that the t-test with substitution works better than a rank-sum test?

There is no assumption by the rank-sum test that the level of variability in each group need be equal.

12-4 Be consistent in using either 'limit' or 'bound' for one-sided bounds. So the first sentence should delete the "interval or", and just say "a one-sided prediction limit is recommended" if it is only a one-sided bound being discussed. Your footnote on page 12-1 should go into the main text, defining the use of interval versus limit. And some places 'limit' is used, while in other 'bound' is used, both for one-sided thresholds.

Stay consistent. So the second paragraph should begin "The 'primer' on prediction limits presented below…" if only one-sided limits are being discussed. Don't use interval anywhere in this chapter unless it specifically refers to a two-sided context. The heading of 12.2.1 should use 'limits' instead of 'intervals', for example. Or 'bounds'.

The third paragraph can be delayed until Chapter 13. It is out of context here.

12-5 "highly non-detect data"? Bad English. How about "data with a high proportion of nondetects"?

12-16 I don't think the power comparisons, with equations, of the two procedures will be used by the target audience. Isn't this a bit too complex? Instead, some worked out examples presented as tables would better allow the audience to see how many samples are needed to get which power level for 10%, 20%, 50% increases in concentration.

12-31 What happens to your intervals if the detection limit is substituted for all nondetects? When 0 is substituted? If the conclusions change, your arbitrary choice of one-half will likely lead to lawsuits. There is no justification for that choice.

13-9 How does figure 13-1 mesh with substitution of one-half DL? If this were reality, a value much lower than one-half would best represent these data.

13-10 Power has been discussed several times previously. This section should be shortened greatly.

Sec. 13-3 This is a long section that goes over reasons why methods recommended in past documents are not used here. Is it necessary is this guidance to do this? This chapter is very long, repeats much of the power discussions of previous chapters, and then discusses methods not recommended for use. I suggest shortening this chapter considerably by focusing on methods that ARE recommended. The other discussion can be put on the web somewhere. Start with section 13-5 and concentrate on what should be used.

14-17 Stated here is a cutoff that control charts are not recommended when more than 25% of measurements are nondetects. This limit wasn't stated in the control chart section. How to construct a control chart with 20% nondetects is not given in that section. The guidance in the control chart section is essentially "don't use control charts when there are nondetects". Given the difficulties in computing a sum with nondetects present, the CUSUM control chart seems particularly difficult to apply with nondetects. All of this being said, its best to delete this 25% limit that suddenly appears here. Stick with the last sentence – its hard to see how to apply CUSUM charts to data with nondetects.

Sec 14.4.1 Your long re-derivation of Spearman's test must be from some reference, perhaps Lehman (1975)? Then don't repeat here. It would seem that

presenting this test in terms of Spearman's rho correlation coefficient makes more sense to the practitioner. Rather than an abstract statistic D, the test statistic becomes the rank correlation coefficient, on the same scale as Pearson's so easily grasped. For the sodium data, Spearman's rho equals 0.758, with the same p-value of 0.011 (normal approx.). Use this version instead of the D statistic. This is after all just a test of the correlation between concentration and time.

Sec. 14.4.2 Similarly, the Mann Kendall test is just applying Kendall's tau correlation coefficient to an x variable that is time. Mann got his name on a test very cheaply. He just said to use Kendall's tau with an x of time. As with Spearman's, presenting this as a test for significance of a correlation using a correlation coefficient will make more sense to a practitioner. Nondetects where there is one detection limit, to be more specific than your statement, can be handled without modification. For more than one detection limit, the Kendall's procedure can easily be modified. See Helsel (2005) for info about this situation, which will be all too common for your readers. Usefulness with more than one detection limit is a large advantage of Kendall's coefficient over Spearman's.

Again on page 14-20 you state that the approximation for Kendall's is good for n>20. This is incorrect, it is excellent for n>10. Spearman's rho requires at least an n of 20 or more for the approximation to work well. However, the exact distribution of Spearman's was only worked out up to n=10 as of 1955, and only is found to n=10 in textbooks. It is not because the approximation is OK at that low sample size. To quote Kendall (1955): "The distribution [of Spearman's rho] tends to normality more slowly than that of tau, and an intermediate form is necessary to bridge the gap between the values for n=11 and the (rather doubtful) point at which the normal approximation is safe;". Of course, as with all scientists, Kendall was not shy in trumpeting the advantages of his statistic over others. But the fact remains, tau has a much better approximation than rho.

It would be good to mention which stat software will calculate rho and tau (most any will do rho if the user first ranks the data. Then Pearson's r is just computed on the ranks). SAS computes tau. There are free and low cost Excel macros on the web that will do both, as well as other nonparametric tests. Or maybe list a an EPA web page (so that it you can update it) that lists which methods are performed by which software, commercial and free/shareware.

14.4.3 Sen's slope has a natural connection with Kendall's tau. It is the slope which, when subtracted from the data, produces a tau of zero. It is computed in an analogous way to Kendall's tau. It has no natural

connection to Spearman's rho. Therefore, while it "can be used in conjunction" with rho, it makes much more sense to do so with tau.

You infer that Sen's slope is easily computed with nondetects, but do not explain how to do it, nor give an example. Since the slope does involve the magnitude, what number do you use for a <1 when comparing it to a later value of 5? It certainly between 5-1=4, and 5-0=5, but where? Helsel (2005) gives the procedure, which is valid for more than one detection limit.

15-12 Falsification of data is never a "reasonable compromise". If I were to 'impute' a value of 0.5 and get a different result when comparing to my confidence interval, which one of us is correct? It would be better to use Kaplan-Meier or Helsel-Cohn for computing means and standard deviations with censored data.

15-14 The estimate of standard deviation for a lognormal distribution is still imprecise for many more than 4-8 observations. Performing Land's procedure for these small data sets is asking for trouble. Either forget Land's method, or put a much higher requirement for data on using it.

15-19 Your example for a NP confidence interval on a percentile works fine for one detection limit. What if in your example, one of the <5s was a <10? How would you rank a <10 versus an 8? There is the simple solution, setting all below <10 as that value. So one end and perhaps more of the interval would be simply <10. More sophisticated solutions are in Helsel (2005).

15-23 As you state, most existing regs are based on a mean. Why then devote all this time to intervals around a median? Can you give examples of when the median is the appropriate statistic? Without guidance, when would a responsible party conclude that using a median is appropriate?

15-31 Here you use the more standard term "confidence bound" for a one-sided threshold, rather than "limit" used earlier. Be consistent throughout, so pick one and use it, but not both. I recommend "bound".

This section, like many others, badly needs references. You did not come up with these binomial formulae from scratch! Far too much of this document ignores all the work that has gone on before by others.

Sec. 15.8.1 Helsel and Hirsch (2002) present an easy way to compute a confidence interval for the Sen trend slope. This can be expressed for any x value, and so provide a confidence interval for the <u>median</u> concentration for a given time x. The Sen slope is a median slope. You jump here to using a parametric linear regression estimate and CI on the mean, when prior you

used a Sen estimate of the median to define trend. This is very confusing, and unnecessary. Your definition and recommendation for a trend test has changed. Either be consistent and compute a CI on the Sen median, or give a better reason for why you are switching to a parametric method for the mean here. There is a very straightforward method for computing a CI for the NP median line, so the reason must be something other than that.

15-36 Substitution of one-half the DL, as you have recommended throughout the guidance, would do just what you warn against here, produce a trend that has nothing to do with sample concentrations. You recognize the problem here – it is just as virulent in two-group comparisons or other situations as it is here. There are good methods for doing regression with nondetects. But you have not addressed them in the Guidance.

Sec. 16.5.2 You state that up to 75% of data can be substituted using modified Aitchisons, otherwise known as subbing in one-half the DL, "assuming the non-detects represent a discrete population distinct from the quantified measurements"? Substitution has crept up to 75% here! The qualification can be rephrased as: "assuming that all nondetects are actually at one-half DL, subbing in one-half DL for them works well". This is unacceptable. It is discouraging to see bad ideas camouflaged with semi-technical jargon.

This section covers two very important topics discussed in detail throughout the guidance, and tries to summarize, I suppose. It fails. The generalizations are not helpful, and in the case of nondetects, are inaccurate. This section adds nothing to the guidance.

Chap. 16 This chapter restates much of what has come before. Little is added. It should be blended in with Chapter 15, to shorten the guidance document.

Sec. 17.3.3 Here Cohen's method can be used with up to half non-detects? In other places it was no more than 20%. No reason for the change, or quantitative justification is given. Recommendations here and in Chapter 16 seem to be musings, with little to back them up. Statements such as "often the best tack is…" would be best ignored by the reader. Unfortunately they won't be. Delete the recommendations for handling nondetects in Chapters 16 and 17.

Chap. 17 This chapter repeats much of what goes before, modifying it only to look for decreases rather than increases. It can be radically shortened, or better, combined with chapters 15 and 16 to eliminate the frequent replication.

Appendix B Simulation study results are very dependent on the characteristics of the data generated to do the work. Different characteristics of the data will result in different procedures being 'the best'. Another cause of

differences between recommendations of 'best' methods in different studies is the set of methods considered for use.

Previous studies that attempted to characterize water quality data found that substitutions of zero, one-half and the detection limit did not work well. There are several such studies listed in Helsel (2005). Methods that did work well for estimating descriptive statistics included the Helsel-Cohn probability plot method, and MLE.

This Appendix simulation did not evaluate either of the above methods. So it cannot determine whether the methods recommended here would outperform, or be worse than, the methods found to work well by others. Based on the work of Gilliom and Helsel (1986), Helsel and Cohn (1988), Hinton (1993), and Shumway et al. (2002), and other studies cited in Chapter 6 of Helsel (2005), a good guess would be that the methods recommended in the Guidance are far worse than those found to perform well in the past by others.

This simulation study started by generating sets of data whose values below the QL were actually all zero. It found that your method which assumed all nondetects were zero, performed the best of the three methods evaluated in estimating the mean. No surprise there – data matched to method. Substitution methods have been found by others to underestimate the standard deviation, and indeed the confidence interval lengths for the methods in this study were cited as being too short. No surprise there. Too bad better methods were not evaluated.

Then it simulated data uniformly spread between zero and the QL. This may not be similar to water quality data (others have not used this uniform distribution), but matches exactly with plugging in one-half the QL to estimate the mean. No surprise then that assuming all nondetects were right at the center of that interval at 1/2QL performed best for estimating the mean. The simulation did not consider the practical issue of QLs being set differently by different labs, as being affected by interferences, and other causes that would be better simulated by randomly increasing the QL value by 3 or more for a third or so of the data. Then substituting one-half of that QL. It did not include a simulation of the recommended method in Chapter 10 of insider censoring – biasing the substitution by taking the lowest values below the DL and calling them <QL. So in summary it did not use realistic issues about the QL value before using one-half that value in the substitution process. The method was found to work well for data simulated to match the method.

The third simulation "model" was to use a normal distribution for the entire distribution, detects and nondetects. What this says beyond that found by Gleit in 1985 is unknown. Cohen's MLE method targeted to a

normal distribution came out best of the three methods tried. No surprise there.

In all three cases the characteristics of the generated data were unrealistically matched to fit the assumptions of the method to be tested. So the corresponding method was found to work well. Methods found to be far better than these in other studies were not considered. But the results of these simulations is cited by the Guidance to support the use of substitution or Cohen's method, instead of alternative, better methods. This is just not scientifically defensible.

References used in the review:
Gleit (1985) ES&T 19, 1201.
Gilliom and Helsel (1986) Water Resources Research 22, 135.
Helsel (2005) Nondetects and Data Analysis. Wiley, New York.
Kendall (1955) Rank Correlation Methods. Griffin and Company, London.
Millard and Neerchal (2001) Environmental Statistics with S-Plus. CRC, Boca Raton, FL.
Shumway et al. (2002) ES&T 36, 3345.

EPA Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities--Unified Guidance
Review Comments by Jim Loftis
June 22, 2005

General

1. Does the Unified Guidance meet the stated objectives as a whole?

a. Does the UG effectively address the performance standards set forth in federal regulations and provide an effective framework for applying statistical methods for groundwater monitoring?

The answer is definitely yes, as far the RCRA regulations are stated and discussed within the Guidance document. I did not conduct a separate review of federal regulations. The background on and discussion of regulatory issues within the document is very good.

b. Is the Guidance presented in a manner that will be accessible to groundwater professionals with a limited background in statistics.

The Guidance is clearly targeted to the stated audience and makes a strong effort to reach that audience as effectively as possible. Most of the Guidance is very clearly written, and most of the material should be accessible to scientists and engineers who work in the field of groundwater—assuming that they have mastered the concepts of an undergraduate upper-division course in statistics, and that they spend a large amount of time in studying the Unified Guidance document and, to a lesser extent, other reference materials. However, these qualifying assumptions should not be taken lightly. The material is complex and can be very confusing to professionals who do not have a great deal of experience in this field or who deal with statistics only on an infrequent basis.

There are some sections that are particularly problematic for use by the target audience, and I have noted these with explanations in the detailed comments that follow.

2. Overall, is the document well organized and cross-referenced in a manner that will help users apply the Guidance?

Yes, the Guidance is well organized and cross-referenced for user convenience. My most significant concern regarding organization has to do with the flow charts, tabular summaries of methods, and Figure 5-1. I am undecided about whether these summary materials should

come at the beginning or end of the Guidance document. The reader who is not familiar with RCRA statistical analysis will find the flow charts in particular to be very frustrating.

Since the flowcharts are designed to be used frequently, it might best to leave them at the beginning of the document but to at least put a statement in front of this section recommending that new users of Guidance read the relevant detailed sections before spending too much time on the flowcharts.

3. Questions a, b, c, regarding each of the five major sections of the document.

Please note that my answers to questions #4, #5, and #6 regarding technical validity (#4), strengths and weaknesses (#5), and additional methods (#6) are included here under my responses to a, b, and c for each major document section and chapter. References are cited where appropriate and are listed in a reference section at the end of the review.

Section I--Regulatory and Statistical Overview Chapters 1-5.

a. Does this section meet the stated objectives from the charge?

I believe that the objectives are accomplished. The regulatory background and requirements are well covered. The historical discussion is very good and should be very helpful to those who have been working in the field for some time. I have recently met agency personnel who are still using CABFT and who will really appreciate this section.

The basic concepts and objectives of ground water monitoring are well described.

Page 3-24 does a good job of pointing out that ultimate determination of whether or not an actual release from a facility has occurred should involve expert assessment of site hydrogeology. Unfortunately this point is not adequately reinforced throughout the document. At several points, the document says that advice from a professional statistician may be needed, but in most of these cases a physical interpretation of the data will be needed as well.

Chapter 3 does a good job of distinguishing between a statistically significant change in ground water quality and a practically significant change in ground water quality, and the rest of the document is consistent in this regard.

b. Does this section cover an appropriate range of topics? Are there topics that are missing or should be explained in greater detail?

This section introduces most of the topics that are discussed later in more detail. The section does a good job of placing the topics into a regulatory and historical perspective. The level of detail is generally adequate for the purposes of this section. However, there are two specific topics that should be discussed more thoroughly, either here or later in the document.

One topic or at least point that is missing from Chapter 3, section 3.7, in the discussion of interwell vs. intrawell tests, is that these two approaches are almost never directly comparable because they are not testing the same thing, either statistically or physically. The actual populations being sampled (set of all possible ground water samples) are different in each case. Therefore appropriate statistical models and hypotheses that derive from them are actually different. Interwell testing is looking for a difference over time, and intrawell testing is looking for a difference over space. While I realize that this is a statistical Guidance document, it seems that the decision of which approach to use is at least as much of a hydrogeological one as a statistical one.

c. Is the material in this section organized in a clear and concise manner?

Some of the material in Chapter 4 would be helpful in understanding the statistical concepts that are presented in Chapter 3; but the reader of Chapter 4 also benefits from the ground water monitoring perspective established in Chapter 3. On balance, I believe that the organization is OK as is.

There are a few places that lack clarity or correctness:

Page 2-19. "In fact, use of the remaining successive downgradient samples might not necessarily improve monitoring performance." This is only true in a very restrictive sense and can confuse the reader at this early stage. "eight annual samples per well" should be worded as "eight samples per year per well".

Page 3-22. "By doing this, the background sample size can be increased significantly even over a short period of time without necessarily violating the assumption of statistical independence." This statement may be true in some sense, but is extremely misleading since spatial and temporal variability in ground water are often, probably usually, quite different , and sampling over space is not equivalent to sampling over time.

Page 3-26. The figure shows conductance in units of micro-moles/ cm^3. There are no such units of conductance. Perhaps micro-mhos/cm were the intended units.

Page 3-30. Here and later, "non-parametric intrawell testing is not generally recommended by the UG." However, the recommended alternative, interwell testing, is not appropriate when there is significant spatial variability. Intrawell testing with retesting will be the best alternative in this case, and if the distribution is not normal, then the testing must be nonparametric.

Page 3-31. The use of a pooled variance to enlarge the background sample size when there is spatial variability among multiple background wells is first mentioned here. I believe that the later use of this concept is incorrect. While the estimate of the variance may be improved, the estimate of the mean is not.

Page 4-3. Equation 1.1 should be numbered as equation 4.1.

Pages 5-28 and 5-30. The detects-only probability plot and Modified-Aitchison Adjustment are based on assumptions that non-detects come from a different statistical distribution than detects. While it is fairly common in environmental monitoring to have observed data that are well described by a mixture of two distributions, it is suggested here that detects and non-detects could, in some cases, be generated by two different statistical processes. This is definitely not the case with the ground water processes that we are dealing with. All of the observations come from a single physical population of ground water sampling units. Nondetects result from censoring the data at some limit, not from a different process. At the least, extreme caution should be used when applying a statistical model that is contrary to the physical process of interest.

Page 5-48. Line 5 should read "—True geometric mean concentration at the compliance point ..."

In both chapter 5 and chapter 15, the distinction between the lognormal geometric mean and the lognormal arithmetic mean is clearly drawn. However, the Guidance is weak on how to decide which of the two statistics should be used in a particular case. My suggestion would be to recommend the use of lognormal geometric means as a default. The default would be overridden if the lognormal arithmetic mean were required by a groundwater standard, for example a HAL that is based on long-term average exposure. The lognormal geometric mean is the appropriate default because it is a better measure of central tendency for lognormal distributions than is the arithmetic mean, and the associated statistical analyses (estimation, confidence intervals) are easier to perform.

Page 5-55. Note that a linear trend in log transformed data would correspond to an exponential trend in the untransformed data.

Section II—Statistical Check and Adjustments, Chapters 6-10

a. Does this section meet the stated objectives from the charge?

This section does generally meet its stated objectives from the charge. The discussion and explanation of methods, including examples are clear.

However, there are several issues that deserve further attention as described below.

b. Does this section cover an appropriate range of topics? Are there topics that are missing or should be explained in greater detail?

There are several topics that need further discussion.

Non-normality:
Chapter 6 stays the course set in previous Guidance with regard to an assumption of normality. That is to say that a normal distribution is assumed unless a test for normality can be employed to reject this assumption. Via the simulation study of normal vs. lognormal prediction limits, a convincing argument is made for a default assumption of normality when using parametric prediction limits with retests, and I accept the argument for that particular case. Since parametric prediction intervals with retests are really the cornerstone of the Guidance, I think that the level of attention given this topic, including the simulation study, is appropriate.

However, I am concerned that the reader will get the mistaken impression that a default assumption of normality is appropriate in more general applications. As pointed out in the Guidance, tests for normality are not very powerful. What is not mentioned is that departures from normality that are too small to be detected by tests for normality can have a marked effect on both the significance level and power of statistical tests other than parametric prediction intervals with retests. This is phenomenon, which is well known among environmental statisticians and water quality hydrologists, is somewhat glossed over in the Guidance.

I do not believe that taking the safer view, that most water quality data are not normal unless "proven" to be so would necessarily result in a different set of recommendations for statistical methods than is currently contained in the Guidance. However, for methods other than prediction limits with retests, I believe that future work should include a careful analysis of the effect of making incorrect decisions on the form of the distribution. This work should include the simulation of both log-normal and "slightly" non-normal data that are subjected to a test for normality and then to either a parametric or competing nonparametric method based on the result of the test for normality. Then the power and significance level of the overall analysis can be evaluated.

Page 6-19. Section 6.3. The discussion of coefficient of skewness should include or at least reference critical values for testing the significance of the skewness coefficient. These are available in Snedecor and Cochran (1980).

Page 6-34. As I mention above, more Guidance is needed on the use of a geometric mean or median as a centrality parameter.

Equal-variance assumption:
It is appropriate to have a chapter on testing for equality of variance, and Chapter 7 fills the bill nicely. However, too much is made of this issue throughout the document. It is stated again and again that various statistical tests, such as Student's t and ANOVA assume equal variances and cannot be used on data for which this assumption is violated. However, with regard to validity of the test (apart from power) this is simply not true. For assessing the validity of the test, the assumption of equal variances applies to the null hypothesis only. When the null hypothesis is true, one must have equal variances in the populations being compared in order to achieve the nominal significance level and a valid test. As long as the test achieves its nominal significance level when the null hypothesis is true, it is valid by definition. Validity does not depend on the types of alternatives for which the test may be applied. This is a fundamental concept that is not at all clear in the Guidance.

Of course, power is important as well, and it may well be that that Student's t and other methods are not the most powerful choices for alternatives (in our case changes in ground water quality) that include a change in variance as well as a change in mean. In that case, then the use of Welsh's t and other methods that are more powerful under conditions of changing variance are entirely appropriate. But I think that it is important to separate the issues of statistical validity (which related to significance level) and power. For most of the cases of concern here, the assumption of equal variance will apply when the null hypothesis is true since the physical interpretation of the null hypothesis is that the observations from all wells or from both background and current samples come from the same physical population of potential ground water samples. If the variances are different, then the populations are not the same.

Page 8-3. "The Unified Guidance does not necessarily recommend automatic screening of the background data for statistical outliers". I don't understand why not. It seems that it should always be done.

Page 8-9. Section 8.4.2. If Dixon's test is run prior to screening for outliers, should it be run with or without the suspected outlier(s)? My recommendation would be to run it without the suspected outliers.

Page 9-16. As noted above, for validity of the test, the equal-variance assumption applies to the null hypothesis only. ANOVA is still valid as long as the variances are equal when the means are equal. The more important issue is loss of power when the variances are not equal, and this is well discussed on page 9-16.

Page 9-21. There is no derivation or reference given for the use of ANOVA to increase the degrees of freedom as explained on page 9-21, in Example 9-3, and at several points later in the Unified Guidance. I do not believe this approach is correct.

It may well be that ANOVA will provide a better estimate of the variance than will a single well, if the variance is in fact constant across all the wells—which is not likely if the means are different. However, the equation for the parametric prediction limit, equation 9.8, is based on an estimate of the variance of the sample mean, ybar. And the variance of the

sample mean is σ/ root n, where n is the number of observations used to estimate the sample mean, i.e. the number of observations at the single well of interest.

This is the best we can do, even if σ is estimated perfectly, with infinite degrees of freedom. Since we are assuming independence across wells and different means among the wells, the additional observations used in the ANOVA provide no additional information about the sample mean, only about the estimated variance. So we might be able to increase the degrees of freedom associated with the t-statistic, but not in the 1/n term under the radical as shown in equation 9.9. I think that this is a recurring mistake in the Guidance.

Temporal and spatial patterns:
Page 9.24. Section 9.4 is very unclear. The term "temporal effect" is very vague and subject to misinterpretation or misapplication. The following indented section describes some of the major problems.

> Temporal correlation will not necessarily show up in parallel across all of the wells. It will show up as short-term trends in a given well. A parallel pattern will result only if there is a common seasonal pattern and/or spatial correlation among the wells. It is likely that there would be both temporal and spatial correlation among background wells at a given site, and there could be a common seasonal pattern as well. However, temporal correlation, seasonality, and spatial correlation are three different things, but these concepts are all mixed together in this section.
>
> Furtheremore, the ANOVA test for temporal effect is not specifically a test for any of these three phenomena. It is simply a test for equality of means across time. Inequality of means across time could result from any of those three factors or many others. As an aside, if we really wanted to test for a common temporal effect on the means, we really should use a two-way test here since observations are taken at each well during each time period.
>
> It seems to me that if there is a temporal effect, we really need to know what is causing it or at least how to describe it properly in order to account for it in the statistical analysis. For example if there is a seasonal pattern, then we need to de-seasonalize the data or use a proper seasonal test, either of which will effectively reduce the variance, not reduce the degrees of freedom.
>
> If there is spatial correlation that extends among the background and compliance wells, then the background wells actually present a better picture of the regional mean than if they were independent. This is the principle behind kriging. The effect will not be to reduce the degrees of freedom. In fact it will be just the opposite unless the upgradient wells are correlated with each other but not with the downgradient well. And if that is the case, then it is doubtful that the upgradient wells actually represent background conditions.
>
> A similar argument extends to temporal correlation, for which the Rank Von Neumann Ratio test of section 9.4.2 is indeed appropriate. If a well or group of wells exhibits serial

correlation, the effect will be to improve the estimate of the mean for the period of record, but worsen it for the long term. Again, this is the principle behind kriging. For immediate comparison with downgradient wells, reducing the degrees of freedom is probably going the wrong direction. There is no clear guidance in section 9.4.2 regarding what one should do if one finds serial correlation using the Rank Von Neumann test, and the question of what to do is not a simple one. The correct course of action when using serially correlated data depends on whether one is making short-term or long-term comparisons to the background data (see Loftis, et. al. 1991).

However, I believe that the safest and simplest course of action would be to recommend that when serial and/or spatial correlation is obvious, downgradient wells should be compared with upgradient wells over the same time period with no adjustment to the degrees of freedom, assuming that all the wells exhibit the same correlation structure. If seasonality is present, it must be accounted for separately.

This section should also note that the Rank Von Neumann test will detect seasonality, which is different from serial correlation and requires different treatment as noted above. Therefore, seasonality should be removed before application of the Rank Von Neumann test if one is really interested in detecting serial correlation.

Contrast example 9-6 with example 9-3. In 9-3 ANOVA was used to increase the degrees of freedom, but the wells were assumed to be independent with different means. Thus the correction was not appropriate. Here the wells are assumed to be spatially and/or temporally correlated, and ANOVA is used to reduce the degrees of freedom, which is also inappropriate. In both cases, the problems stem largely from the lack of a precise statement of the statistical model being assumed. While I understand the need to limit statistical complexity as much as possible, I think that the effort to achieve simplicity has led to some mistakes in these two cases.

Pages 9-37 to 9–40. The discussion of seasonality is good. To simplify the discussion, though, I do not think that one needs to worry about seasonal cycles of length other than one year; and if such cycles are suspected, one should consult a hydrogeologist, not just a professional statistician. This last comment applies generally, not just here. Most professional statisticians will not have the background and experience to handle ground water statistics effectively without help from hydrogeologists.

Censored data:

Chapter 10 in general. There is no mention of the fact that information is always lost when analytical results are censored. The statistical procedures that are described in this Guidance will provide better results if un-censored concentration data from the laboratory, including negative concentrations, are used instead of censored data (Porter, et al., 1988). While there is still great resistance from laboratories to providing uncensored data, this should at least be mentioned as an option for improving statistical performance of ground water monitoring programs.

Page 10-17. I think that the physical basis for a model in which non-detects are assumed to come from a separate distribution than the detects needs to be explained in greater detail. The Guidance suggests that use of this model should be tied to the physical situation, but I can't see many, if any, physical situations that would produce such a result unless contamination "comes and goes." And if contamination is ephemeral, then it will indeed be very difficult to deal with. See my comment above about pages 5-28 to 5-30.

Page 10-28. I do not understand the statement "Note, however, that highly diluted samples sometimes have reporting limits that far exceed the maximum observed detected value." I assume that what this means is that if you dilute a sample by a factor of 10, the effective reporting limit is ten times the reporting limit of the analytical method. But I don't see why that has any relevance. The only acceptable observations that are censored should be ones that are not diluted. If you get a nondetect or trace result on a diluted sample, you have diluted it too much, and you have a laboratory error. Thus the resulting data point should not be used in the statistical analysis.

Page 10-29. The tabular source of the kappa values for eq 10.11 is not given.

Page 10-31. How do you know for sure that one result is mistaken vs. another? This is not a simulation in which the true answer is known. All you can say is that the data appear to match one model better than another.

c. Is the material in this section organized in a clear and concise manner?

Figures 6-1 and 6-2 are not really figures. They are tables. I don't think that these tables really belong in the main text since their function is simply to support the case for a default assumption of normality. They might serve just as well in an appendix.

Page 9-20. Step 8 in the example does not belong here. The use of ANOVA to obtain an improved estimate of the variance is not discussed until the next section, 9.3.3, and it has nothing to do with this example, which is about testing for equality of means.

Section III—Methods for Detection Monitoring, Chapters 11-14

a. Does this section meet the stated objectives from the charge?

This section does generally meet its stated objectives from the charge. The discussion and explanation of methods, including examples are generally clear. However, there are several issues that deserve further attention as described below in part b.

b. Does this section cover an appropriate range of topics? Are there topics that are missing or should be explained in greater detail?

There are several topics that are either not appropriate or deserve fuller explanation as noted below.

Page 11-8. It should be noted that in any real situation, there will be a difference in means between two wells or two periods in time. Even if it is very small and practically signficant, the difference will not be zero.

Page 11-13. This discussion on power, involving the non-central t distribution is good, but I think that it is beyond the easy grasp of the intended audience. I don't have a good solution for this, but I wonder if this and other Guidance recommendations regarding power calculations (that require more than simple tables or graphs) could not be simplified greatly by using rules-of-thumb that are developed from the extensive power calculations that have already been done or could easily be done by experts. This is perhaps an area for further research. As things stand, I fear that we are asking too much of our intended audience.

Page 12-21. Footnote 4 says " Note that in select cases, no retest will be required ..." There needs to be more information about what the "select cases" are.

Page 12-22. In this example there is no statistically significant difference at the 0.01 level, but clearly there is a difference between the two groups. This is a great opportunity for some additional discussion to reinforce the distinction between a practically significant and statistically significant difference and about the effect of doing comparisons at the 0.01 level vs. larger levels of significance.

Page 12-27. Given the obvious problems with Poisson prediction limits pointed out in Loftis, et al. (1999), it seems odd that they continue to be recommended. Since Poisson prediction limits do not play a major role in the Unified Guidance, it is perhaps not a huge issue in most applications. But several conceptual and theoretical problems could be eliminated and a simpler Guidance could be achieved by eliminating this method. Very little would be lost.

To be more blunt, no scientist should ignore units in his or her analysis, and one cannot use this method without doing just that. Equation 12. 17 has units of concentration in the first and third terms and no units in the middle term. The suggested approach of regarding concentrations as "counts" is arbitrary and unsound. As soon as a laboratory method improves, a count of 1 will become a measurement of 1.2 or 1.24. You can count fish, but you cannot count VOCs.

The idea of rescaling is also arbitrary, and the result could depend on how one rescales the data. One of the main reasons for a statistical approach to data analysis is to reduce or eliminate the need for arbitrary or subjective decisions, so we should rely on methods that are

more fundamentally sound and less arbitary than Poisson prediction limits. Please at least note the problems, cite the reference, and let the user make a fully informed decision.

Page 13-1. I would avoid the use of the term "dirty" which implies lack of suitability for an intended use. I would suggest "impacted" instead.

Sections 13.2.1 and 13.4. Though the effect of correlation among wells and among constituents is mentioned, I think that a bit more explanation might be helpful.

The equations presented for FWFPR are really a worse-case scenario since they assume independence. At the other extreme is an imaginary case where all of the variables and wells are perfectly correlated so that you get identical information from all of them. In that case the FWFPR is just alpha. The real situation is somewhere in between, but in practical applications we assume independence to be conservative. To take the spatial correlation structure into account in determining the FWFPR would be extremely difficult.

Page 13-18. In example 13-2, the explanation of the simulation procedure is not really clear. However, as I read it, step 1 is not really used until after the power curve is developed. It seems that the power curve is developed for standardized measurements and then interpreted after the fact for a population with a given CV. But the directions are not clear.

I am also concerned that this level of analysis, involving simulation, is beyond the easy grasp of the intended audience.

Page 13-32. The discussion of the FWFPR when both intrawell and interwell testing are used is extremely confusing. At the bottom of the page, we read that the significance level for each of c interwell tests is α/c and is $\alpha/(r-c)$ for each of (r-c) intrawell tests. It seems to me that this leads to a FWFPR of 2α. This needs to be cleared up.

Page 13-33. Step 2. Within the example itself there is no indication of where n^* and s^* come from. In any case, as stated earlier, I believe that the ANOVA approach to correcting for a temporal factor is incorrect.

Page 13-42 and 13-43. We have the same problem as described earlier regarding sample size and degrees of freedom. The value of κ should depend not only on how well S estimates the true standard deviation, but also on how well xbar estimates the true mean. The degrees of freedom used in the example will overstate the latter, as I have explained previously.

Page 13-49 and page 13-52. These procedures have the same problem as above with adjusting for temporal effect.

Page 13-56. The URL for the Optimal Values Rank calculator is missing.

Page 13-60. The paragraph starting with "If absolutely called for ..." is not clear. It is unclear what "absolutely called for" means, and the sentence starting with "If not,..." is unclear as well. It should probably read "If none of the nonparametric prediction limits is

sufficiently powerful, then ..." The last sentence is also unclear. Probably what is intended is that the user should select from the various alternatives in such a way that the "best" compromise between FWFPR and power is achieved, but it really says that adequate power should be achievable even if a reasonable FWFPR is not. Is that what is intended?

Page 13-71. Computed power curves should be presented so that the user who attempts to implement this approach could check his or her results. Again I am concerned about suggesting that the target audience rely on simulations for their power analyses

Page 13-78. The pooled degrees of freedom from ANOVA is used again here.

Page14-19. I do not see why the Guidance needs to include both Spearman's and the Mann-Kendall test for trend. It would seem that they are redundant, and the result is unnecessary complexity of Guidance. I would include just the Mann-Kendall test and perhaps a reference for Spearman's as a legitimate alternative.

A powerful alternative that is much easier to implement if one has only a spreadsheet is linear regression on ranks. See Conover (1980) and Taylor and Loftis (1989). This could be a useful addition to the Guidance for some practitioners.

c. Is the material in this section organized in a clear and concise manner?

The organization of this section is good. The problems with clarity can be addressed with additional detail as described above in part b.

Section IV—Methods for Compliance Monitoring and Corrective Action, Chapters 15-17

a. Does this section meet the stated objectives from the charge?

This section as a whole meets its stated objectives, but marginally. There are two main problems with this section. First the material is very complex, and the discussion is difficult to follow in several places, particularly those dealing with statistical power. Chapter 16, section 4 is particularly troublesome, and in my opinion, is not usable by the intended audience. Second, this section, and in particular Chapter 16, section 4, contains several important errors and inconsistencies.

Chapters 15 and 17 accomplish their objectives, but there are several problems with each chapter, as noted in the detailed comments below.

b. Does this section cover an appropriate range of topics? Are there topics that are missing or should be explained in greater detail?

There are several topics that are either not appropriate or deserve fuller explanation as noted below.

Page 15-19. This is a bit picky, but it seems to me that in equations 15.8 and 15.10 (and 15.17 and 15.19) the term U^*-1 should be U^*. My reasoning is that the upper tail would be the probability of more than U success, not U or more successes. The lower tail would be the probability of getting fewer than L successes, which is OK as written. But the difference could just be in the way that the confidence interval is defined. Or I could just be wrong. It would be good to cite a reference for the procedure.

Also, I would recommend including an approximate formula for nonparametric confidence intervals on quantiles, appropriate for larger sample sizes, such as equations 11-11 to 11-13 in Gilbert (1987).

Page 15-25. Need to cite a reference for equations 15.11 - 15.14.

Page 15-37. Need to cite a reference for equations 15.24 and 15.25. It would be good to note that the regression coefficients, r, and the MSE can easily be obtained from regression software—spreadsheet or statistics package.

Page 16-7. The examples involving an LC_{50} is probably not a good one to use for illustrating standard based on medians for a couple of reasons. The LC_{50} is not the median of anything that I can think of, and one does not compare median concentrations to an LC_{50} to evaluate compliance. The LC_{50} is a toxic limit that should rarely be exceeded. Standards that are based on exposure to a toxic substance are more likely to be related to the mean than to the median since the mean is a better indicator of total exposure.

One type of standard that is based on the median (of a lognormal distribution) is an upper limit for the geometric mean of a microbial count. But, of course, that is not a relevant example for RCRA facilities. Fixed limits based on the median are really most appropriate when the limit is based on ambient conditions, i.e. background data, and those data are not normally distributed. That should be a fairly common situation.

Page 16-8. There is an extra “a” after “minimum” in line 5.

Page 16-14 In line 3, “constant population variance is assumed, equal to the standard,..” That should be population standard deviation, equal to the standard--not the variance.

There is no description or reference for the algorithm used to compute Figures 16-3 and 16-4 until page 16-17. So equation 16.6 should be moved up to the discussion of those figures, and a reference for the equation should be provided.

Page 16-15 . Provide a reference for equation 16.5. I do not know whether it is correct. I was able to reproduce values of Figure 16-5 (which is really a table) using this equation.

However, the table values in Figure 16-5 in the row for CV=1 do not agree with Figures 16-3 and 16-4 . As I read this section, it seems that they should agree. For example choose 50% power at R=1.5, CV=1, and n=12. Figure 16-3 gives α=roughly 0.05, while Figure 16.5 gives α=0.136. If there is not really an inconsistency, then there is definitely a lack of clarity.

Page 16-17. Provide a reference for Equation 16.6. The right-hand-side of this equation must give values of β not 1- β, otherwise it does not give the correct answer for the no change condition of R=1. I do not know whether it is otherwise correct. After making this correction, I was able to reproduce values of Figures 16-3 and 16-4 using equation 16.6. Appendix Tables 16-1 and 16-2 appear to agree with Figures 16-3 and 16-4.

Page 16-19. Define "compound null hypothesis" both here and back in chapter 4 where the concept of a null hypothesis is first introduced.

Page 16-22, 23. In Example 16-1, I think that it would be good to note that if you were to use Figure 16-5 with the sample CVs of around 0.3, you would obtain an alpha of 0.05 instead of 0.163, and a smaller confidence interval would result.

Page 16-30. Please give a reference for equations 16.7 and 16.8. The Table 16-3 values appear to check out correctly with the equation.

This section is fairly confusing. I would recommend restatement here of how the lower confidence limit on the percentile of interest is to be obtained (in both the parametric and nonparametric case, like the statement in Step 3 of Example 16-2 .) It would make things a lot easer if the approximate formula I mentioned above could be used in the nonparametric case.

Page 16-31. In line 9, I think that the "or less" should be deleted. If p is actually less than p_0, the significance level should be less than alpha. Or you could add "or less" to the next line after "alpha x 100%".

Example 16-2. I would recommend using an example based on ground water quality data instead of pressures of chlorine gas.

Page 17-7. Equation 17-2 needs to be referenced.

Equation 17.2 also appears to be incorrect. I believe that the correct equation is equation 7.8 in Zar (1999), which would be the same as the equation shown without the R in the numerator. The equation 17.2 and thus the corresponding tables 17.1to 17.3 will lead to sample sizes that are too small—by a factor of roughly R^2.

The tables are captioned incorrectly because they are indexed by the amount of decrease in the mean which is 1-R in the notation in the text. The correct caption should read minimum n for a risk ratio of 0.75, 0.5 0, or 0.25 and should define risk ratio. The caption could note that the decrease = 1-R.

Page 17-11. Third line from the bottom. Figure 17-5 should read Figure 17-6. Also in the actual figure, it would be clearer to indicate 90% confidence limits in the legend, since the text suggests that the upper and lower limits are each 95% one-sided limits.

Page 17-16. In Step 10, I don't see a correspondence between the sample sizes discussed and Table 17-2. This does not really matter because I believe that Tables 17-1 through 17-3 are incorrect, as mentioned above.

Near the bottom of the same page, it is stated that the minimum sample sizes in Table 17-1 to 17-3 are based on the conservative assumption that the standard deviation is equal to the clean up standard. There is no such assumption in either equation 17.2 or the correct version of it. If we made such an assumption, there would be no need to specify the CV, because the decrease would be expressed as a multiple of the standard deviation instead of a multiple of the mean, as it is now.

Page 17-19. Insert "at most" in seventh line from the bottom. "... the test will still declare the remediation successful at most αx100% of the time.

Page 17-20. In step 1, leave out "only" in "success will still be ~~only~~ 20%. "
The entries in Table 17-4 appear to check correctly with equation 17.3 However, as noted in the text, the required sample sizes are quite large. I would go farther and admit that the required sample sizes are ridiculous—or at least generally far beyond any reasonable expectations for monitoring. It seems obvious to me that this approach--of requiring that an upper confidence limit on an upper percentile be less than an upper percentile established by a GWPS--would not be feasible in very many cases.

However, I think that it is very valuable to present the calculations and associated table values that demonstrate this fact. I would simply acknowledge the impracticality of this approach in the text rather than saying that as more cleanup data are accumulated, the odds would increase of declaring a remediation effort a success. It would take a long time to get 1,000 or more samples.

Page 17-21, 22. The approach to determining whether a remediation has been successful when the background mean is below the GWPS seems too simplistic, and ignores power altogether. It seems much more logical and consistent with the other analyses of remediation efforts to require that the post remediation mean (or median if the distribution is not normal) be within some stated fraction (say 10%) of the background mean at a given confidence level and power. The appropriate test would be a t-test or Mann-Whitney test depending on the distribution, and the appropriate sample size would be approximated by the method of example 11-2 or more simply by equation 8.22 in Zar (1999).

c. Is the material in this section organized in a clear and concise manner?

This particular section suffers from lack of clarity in several places as noted above.

A recommendation that applies to the entire Guidance but is especially important to this section is to include a numerical example for each equation and to include more detail in all of the examples that use table values. State precisely how to determine which table to use, how to find the table entry (row and column), and what table value is found. If an outside reference or program is needed, for example to obtain a non-central t value, indicate the value obtained for the example problem.

Another general comment is that I believe that the Unified Guidance should not present more ways of accomplishing a task than are necessary. Figures 16-3 and 16-4 (data shown in Tables 16-1 and 16-2) and Figure 16-5 are all designed to solve essentially the same problem. Over and above the fact that they do not agree as I noted above (so there must be an error in one or the other), it would be much simpler to choose a single approach—whichever one is deemed simplest yet adequate for the task. Or if both are really needed, then have a clear statement of when to use one vs. the other, such as " if there are fewer than x data points from which to estimate a background CV, use approach #1, otherwise use approach #2."

Section V—Further research/Guidance needs and appendices.

a. Does this section meet the stated objectives from the charge?

There are not clearly stated objectives for this section in the charge.

I don't see that Appendix A adds much to the Guidance. While the issues listed there are indeed appropriate for further research that could lead to improving the Guidance, I do not think that the discussion is of value to the intended audience of the Guidance.

I don't think that Appendix B is of much value for the intended audience either. I have already stated that I do not know of any physical process that would lead to the mixture models that match the assumptions of Aitchison's Adjustment. I would much prefer to see more discussion of how such situations could and do arise in the real world.

The analysis of the application of Aitchison's to a single continuous distribution is valuable.

The statistical tables are, of course, included for specific needs identified in other sections, and they meet those needs. I have noted some technical problems with the tables above. In general, they need more detail, as explained below in part c.

Appendix E, the Glossary, is good. The definitions are not rigorous, but I think that is OK, given the purpose and audience. The definition of confidence level should possibly be expanded a bit. Perhaps say "degree of confidence, expressed as a probability, of a ..." I would add "significance level" to the glossary and note that the confidence level is 1 minus the significance level.

b. Does this section cover an appropriate range of topics? Are there topics that are missing or should be explained in greater detail?

Given the intended audience, I believe that the most appropriate topics for further research are those related to simplifying the Guidance rather than making it more complex. Research should be directed toward reducing the number of alternative methods for accomplishing a required task and evaluating simplifying assumptions and rules of thumb that can be used to make the Guidance as simple as possible.

Our paper, Loftis, et al. (2001) describes research of this sort. In that study, we did extensive analysis and simulations oriented toward developing simple Guidance for practitioners. One conclusion was that "The detectable changes as calculated from the confidence interval approach are not detectable over half the time because the power is always less than 50%". Another conclusion was that "To increase the power of change detection to a more reasonable value of 80%, the sample sizes need to be at least twice as large as the sample sizes calculated by the confidence interval approach."

Also, repeating a recommendation made earlier, I believe that future work should include a careful analysis of the effect of making incorrect decisions on the form of the distribution . By this I mean the effect on power and significance level when the selection of parametric vs. nonparametric methods is included as a part of the analysis. Prediction limits with retests would not seem to require further study in this regard, however.

Moving now to Appendix B, I would also like to see an analysis of Cohen's method when misapplied to mixed models. The objective would be to determine whether or not we might be able to do without Aitchison's Adjustment altogether, thus making the Guidance simpler and more concise.

c. Is the material in this section organized in a clear and concise manner?

My comments above deal with the relevance of certain topics.

There should be a complete list of the tables included at the beginning of the main text.

Additional clarity is needed in captions and headings of the tables. Each table should stand on its own without need to refer back to the text to see what the table is for and what all of the notation means. There should also be a statement of how the values were obtained, referring to an equation number or simulation description back in the text.

As examples:

The heading for Table 6(5) is indecipherable on its own.

Table 6-6. The heading should define G_i.

.
. (and so on)
.

Table 13-30 should include "Median of order 3" in the heading of each page.

Tables 16-1 and 16-2 are for a confidence interval approach and are based on an assumption of CV=1. Neither of these are stated in the heading.

All of the table headings and captions should be reviewed and improved as needed per the above criteria.

Specific topics

1. Comment on the UG approach to the multiple comparisons problem in detection monitoring.

I believe that the recommended approach, relying primarily on prediction limit retesting strategies, is both reasonable and sound. The approach is able to achieve both reasonable power and significance levels in cases of numerous individual comparisons (well-constituent pairs). While the general approach is not new, I think that the additional material in the Unified Guidance, with greater emphasis and information on power is most appropriate and should prove to be very useful. Chapter 13, though complex, is generally clear enough for the intended audience to implement.

In general, I believe that the increased emphasis on power in this version of Guidance is most appropriate. However, I am fearful that any power calculations that involve simulation or anything (such as the non-central t distribution) that is not included in spreadsheets is probably pushing the limits of the intended audience. I believe that simpler approaches should be recommended whenever possible.

2. Comment on UG approach to multiple comparisons problem in compliance/corrective action monitoring with regard to two major recommendations: (1) a priori power criteria, and (2) aggregation of annual data to enhance power and single-test false positive errors.

I agree with the discussion and conclusions in the UG around the problem of defining or computing a FWFPR in either compliance or corrective action monitoring. As stated on page 16-20, it would seem that power is of more concern than false positives for this type of monitoring, though one should certainly attempt to balance the two. Particularly with regard to Chapter 16 on compliance monitoring, requiring adequate power is essential. Otherwise one could easily end up with small sample sizes and huge confidence intervals that would not be able to detect important violations of GWPSs.

The idea of aggregation of data over time is reasonable. I don't see a simple alternative. But as noted at the top of page 16-21, the resulting analysis has now become a *sequential analysis,* and both the power and significant level of the overall sequential analysis will be different from that of an individual comparison. This deserves more explanation than is currently presented.

In corrective action monitoring, Chapter 17, the proposed definition of achieving a given clean-up limit—entire confidence interval below the limit—imposes a need for power on its own. If this definition is used, the entity performing the clean-up will have a very strong incentive to achieve as high a power as possible in order to demonstrate success as soon as possible. The real problem, as correctly noted in both chapters 16 and 17, is in statistically detecting a difference from the standard when the true mean or percentile of interest is close to the standard, i.e. achieving a reasonable power with a reasonable sample size. And, of course, in chapter 17 the sample size requirements for demonstrating achievement of upper percentile limits can get ridiculous.

My suggestion for improving this section is to emphasize that each case will be different. Thus setting up the criteria for compliance or corrective action and designing the monitoring program and statistical approaches to evaluate compliance or corrective action are complex and highly site-specific analyses. Any such analysis needs to consider the physical/hydrogeological situation, the actual risk levels, and the realities of what type of clean-up can actually be achieved and detected with any reasonable level of investment.

3. Comment on recommendations that represent revision or enhancement to current Guidance.

The revisions and enhancements are listed adequately on pages xvi and xvii of the document. I think the following are most important.

a. Welch's t-test. Apparently this test offers a power advantage over the usual Student's t-test under alternatives that include a change in variance. I do not know whether the advantage is sufficient to justify the extra trouble to implement the test, given that it is not included in spreadsheets or basic statistical software that would always include Student's t. Welch's test is, of course, preferable to CABFT. Sufficient Guidance is presented for its use by the intended audience.

b. There is expanded discussion of interwell v. intrawell methods with guidance for selection of one vs. the other. I believe that this guidance will prove to be quite useful, though I have some specific concerns, expressed above in section-by-section comments. In particular, I believe that there is always some difference among individual wells, and the Unified Guidance is perhaps not quite cautious enough in its recommendations for interwell testing. Furthermore, I believe that the ANOVA approach for increasing the degrees of freedom for intrawell testing when there are multiple upgradient wells with different means is incorrect.

c. The discussion and recommendations regarding confidence intervals are greatly expanded and improved, and the recommendations for the use of confidence intervals on trend lines seems appropriate. The Guidance is sufficiently clear and detailed for successful implementation of these methods by the intended audience. However, in the particular case of nonparametric confidence limits on percentiles, I have recommended above a simpler approach for the case of larger sample sizes.

d. The discussion of temporal patterns, beyond seasonality, is new. I do not believe that the discussion is very clear, however; and I think that the ANOVA approach is not correct as presented. A detailed explanation is included in my comments above.

I also believe that the Kruskal-Wallis test, along with box-and-whisker plots, should be the basic tool for detecting differences among wells or among seasons. This test has largely disappeared from Guidance, and I believe that this is a mistake, given the general lack of normality of water quality data.

4. Are the statistical methods summaries in Chapter 5 useful, and do they provide clear Guidance for potential users?

The summary and flow charts, especially the latter, will probably prove to be useful. The summaries in particular add to the length and repetitive nature of the document, especially since much of the information is presented in both tabular and narrative form. Nevertheless, it is often good to present the same information in multiple forms to match the preferences of different types of users.

I did not find much discussion of Figure 5-1, Choosing an Acceptable Method, on page 5-13, and it is not really clear how to use this figure. But the idea of a one-page summary of method selection is really good.

I did notice one obvious discrepancy between Figure 5-1 and the text. Confidence intervals on percentiles are not checked in the figure for corrective action monitoring. However, this approach is discussed in Chapter 17, and Appendix Table 17-4 is intended for that purpose. Also, I don't follow all of the logic behind the recommendations in Figure 5-1. For example, I don't understand why the various type of prediction limits would not be checked for small facilities. I don't understand why prediction limits should not be checked for long data records. I would think that fixed limit (maximum) should really be fixed limit (upper percentile) since there is really no statistical basis for comparison against a maximum value.

Figure 5-1 should be more carefully thought out and should clearly agree with the flow charts and text description and recommendations.

I found several of the flow charts to be confusing. The main problems I found were the following.

a. There are no Y or Ns on Flowchart 5-3 on page 5-64.
b. On 5-8 on page 5-65, one is directed to use the LCL on the mean (Land's method) rather than on the median. But the discussion later in Guidance seems to give equal weight to CLs on medians for lognormal data—and median is in the title of the flowchart. The same thing happens on page 5-70.
c. In 5-13 on page 5-69, there is no indication of how you decide which way to go from start. Same thing on 5-74, 5-76.
d. In 5-12 on page 5-69, and several other charts, you go to a letter like B or C, which requires you to go back to a previous chart. That gets really confusing. In the circle with the letter (not in the preceding rectangle), it should say " to Flowchart 5-11" etc, and in the flowchart that you reach, the circle should say "from Flowchart 5-12" etc.
e. On 5-12, 5-11, and other charts that are linked, it is not immediately clear from the charts themselves which chart is really the start, even though there is a part 1 and part 2. All of the linkages between charts need to be clarified.

As stated earlier, I am undecided about whether the flow charts should come at the beginning or end of the Guidance document. The reader who is not familiar with RCRA statistical analysis will find them to be very frustrating. It might be good to at least put a statement at the beginning of the flowcharts recommending that new users of Guidance read the relevant sections before spending too much time on the flowcharts.

5. Is the software program for Chapter 13 non-parametric prediction limit testing useful and accurate?

Yes, the program appears to very useful, and appears to be accurate for those calculations that are easily checked. Most of the results are obtained by simulation and are not feasible to check. I did compare the calculator with the appendix tables, though.

The power calculations appear to be fairly consistent with the tables in Guidance, but I did find some differences in the "transition zones" between the optimal and moderate power ranges. For example, for a 1-of-2 plan with w = 10, the appendix table (13-19) transitions from bold to shaded (presumably from optimal to moderate) between n=60 and n=70, while the calculator shows a transition from optimal to moderate between n = 35 and n= 40. The calculator showed no change from n= 60 to n=70, and the table showed no change from n=35 to n=40. The results for significance levels were exactly the same between the tables and calculator in all of these cases.

Since both the tables and calculator rely on simulation for power calculations, it is not feasible to see which result is better, but the difference could be simply that the calculator uses regression approximations of the simulation results. In any case, it would be good to mention the discrepancies in the Guidance.

In addition to such differences in results, there is also a difference between the calculator and Guidance in the terms used to describe power performance "excellent, good" vs. "optimal, moderate", etc. It would be good to use consistent terms.

References:

Conover, 1980. Practical Nonparametric Statistics. John Wiley and Sons.

Gilbert, R. O. 1987. Statistical Methods for Environmental Pollution Monitoring. John Wiley and Sons.

Loftis, J. C., G. B. McBride, and J. C. Ellis. 1991. Considerations of scale in water quality monitoring and data analysis. *Water Resources Bulletin.* 27(2):255-264.

Loftis, J. C., H. K. Iyer, and H. J. Baker. 1999. Rethinking Poisson-based statistics for ground water quality monitoring. *Ground Water* , 37(2): 275-281.

Loftis, J. C., L. H. MacDonald, S. Streett, H. K. Iyer, and K. Bunte. 2001. Detecting cumulative watershed effects, the statistical power of pairing. *Journal of Hydrology* 251:49-64.

Porter, P. Steven, Robert C. Ward, Harry F. Bell. 1988. The detection limit *Environ. Sci. Technol.*; 1988; 22(8); 856-861.

Snedecor, G. W. and W. G. Cochran. 1980. Statistical Methods. Iowa State University Press.

Taylor, C. H. and J. C. Loftis. 1989. Testing for trend in lake and ground water quality time series. *Water Resources Bulletin*. 25(4):715-726.

Zar, J. H. 1999. Biostatistical Analysis. Prentice-Hall.

U.S. EPA Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities Unified Guidance Peer Review Panel

Final Written Comments

William A. Huber
Quantitative Decisions
Merion Station, PA

Introduction

The comments provided here were generated primarily as I read through the draft of the Unified Guidance (UG) from beginning to end. In some cases, my criticisms or questions were addressed by later sections of the UG. I have let these stand, because they reflect the reactions that one might observe in any other reader similarly attempting to follow this guidance by reading it through.

Wherever possible, I have attempted to situate specific comments within the context of the questions set forth in the "Charge to Peer Reviewers." If I had more time, I could effect a better re-organization of these comments. Instead I will apologize here that not all the comments appear in the right places. Nevertheless, I believe I have addressed in detail every one of the questions in the Charge.

This is probably the best place to state that the UG contains some really exemplary and thoughtful material. Many of the background and historical sections in the individual chapters (6 through 17) are clear, well reasoned, and comprehensive. They deserve a wide audience. My comments would be far lengthier were I to take every opportunity to point out such good material. Thus, most of the comments focus on places where there are opportunities for improvement or where I was just plain confused. I do not want the preponderance of such comments to be taken as a reflection of the overall quality of the UG.

There remain some points that do not respond to any specific question in the Charge. These are based on patterns that seemed to develop within the UG as a whole. I will begin my review, then, by raising and briefly discussing these issues. Here and below, most recommendations are underlined. (Some paragraphs are just long lists of recommendations; I did not underline them because it would be redundant and distracting.)

(1) Many states in practice will approve only procedures appearing in guidance somewhere. It would therefore help to provide as comprehensive a list of possible approaches as possible, even those not described in detail. Wherever possible, please

indicate what alternative tests, procedures, or methods might also be applicable, whether or not they are specifically recommended.

(2) Many readers of the UG ultimately will need to write portions of a RCRA permit. The UG can help them in several ways. Provide a section that responds explicitly to all the regulatory requirements: explain the performance requirements and how the guidance meets them. Describe how the permitting process works ideally, with detailed examples of (a) how a test would be described in a permit, (b) what form the justification for that test would take (a separate report?), and (c) how its results would be reported after each monitoring round. Point (b) is especially important, because most RCRA permits have no place where one can document the basis for test selection (including providing results of power calculations). Exactly what kind of formal vehicle does the EPA recommend for documenting power calculations and the other kinds of analyses needed to justify a selection of statistical tests at a RCRA facility?

(3) The UG in its current form could easily be misinterpreted to allow for *post-hoc* test selection. It is not difficult to foresee a spate of permit revisions that state, in essence, that first one test will be tried, then another, then another, and so on, during each monitoring event. Of course many facilities would attempt to use that flexibility to their advantage, running one test after another to snoop for the one that gives the most favorable results. It is important to discuss this possibility and specifically to recommend against it: to the extent possible, RCRA permits should be as clear and specific about which tests will be conducted and what actions will follow from their results.

(4) How will any resamples get incorporated into the monitoring record (especially if they might be used in later tests)? This is a difficult question that the UG does not fully address. In practice, it seems that database treatment of apparently *ad-hoc* measurements varies capriciously from one facility to another: some of them include such results in their databases, others do not, and most do not flag them properly (to show that the samples were obtained contingent upon the results of preliminary statistical tests). This can lead to biases whenever such data are later used for other purposes.

(5) In detection monitoring it actually is fairly common for one or more UG wells consistently to have substantially higher concentrations than one or more compliance wells and/or to exhibit temporal fluctuations. As soon as enough background wells are available, this situation seems inevitably to arise. The concern is that this obvious spatial variation may be dynamically changing, so that the high concentrations observed upgradient sooner or later might appear in downgradient wells. When this is the case, adopting intrawell methods does *not* solve the spatial variability problem; it just overlooks it. Therefore, the framework put forward by the UG might not be as effective or comprehensive as its authors seem to think. At the very least, whenever intrawell tests are adopted, the facility should still routinely compare compliance well data to contemporaneous upgradient data. This ought to be a standard recommendation. In principle, in such a dynamic situation, there cannot be sufficient evidence of a release until concentrations at a compliance well are consistently above the highest

concentrations that one could reasonably project based on both the temporal variability *and the spatial variability* observed at upgradient wells. The obvious solutions (monitor many more background wells; characterize groundwater transport in greater detail) are usually impracticable. I don't have a solution to offer.

(6) Please provide formulas and explicit references for *all* calculations and *all* tables. I have found substantial errors in some of the tables (and no errors in a few of the others), so I have to believe errors may exist in the tables I have not had time to check. You must provide sufficient information to let any reader independently check the accuracy of every calculation and every table. Giving alternative names for the tests would also be useful, because that would facilitate learning more about them through other sources.

(7) Must all facilities actually achieve the targeted 10% annual expected false positive rate? The UG states "the new recommendation of a 10% annual false positive rate puts all facilities and detection monitoring statistical programs on an equal footing … all facilities will then be at equal risk for false positive determinations" [at 13-5]. This seems to suggest that a facility that could achieve acceptable power on all tests might still not receive EPA approval unless its false positive rate were actually 10%. What if it were 1%, or 0.1%, for instance? Would the facility be forced to take a nominally acceptable program and make it worse from the point of view of false positive rates? I hope the EPA will be able to clarify this.

General Topics

1. Does the Unified Guidance meet the stated objectives as a whole? In general,

   a. Does the Unified Guidance effectively address the performance standards set forth in RCRA §264.97(i) and §258.53(h) and provide an effective framework for applying statistical methods for groundwater monitoring?

   It provides an effective framework, but not a comprehensive one. Some of the guidance appears unnecessarily restrictive, by being too specific and envisioning situations that will not universally hold.

   It is clear that much more research needs to be done, and is being done, on sequential monitoring methods (especially control charting), on estimating actual false positive and false negative rates for complex tests that incorporate diagnostic testing and verification sampling, and on interwell tests. By being unnecessarily specific (about things such as alpha levels for diagnostic tests), the UG may make it more difficult for facilities to propose new methods and approaches as they come along in the future.

   I therefore recommend that, wherever possible, the UG emphasize *principles* rather than *procedures*. It can still incorporate some "example" or "illustrative" procedures, such as presented in chapters 11 through 15, but it ought to make clear that other procedures can be acceptable, provided they demonstrably meet RCRA requirements.

   Some of the new principles enunciated in the UG include asking for explicit power calculations, providing quantitative guidelines for minimum power, providing a framework for balancing power and false positive rates, and sets of default assumptions to make about statistical distributions and treatment of nondetects.

   There is a fundamental flaw to this entire framework, though. It is exhibited by the asymmetry in evaluating false positive and false negative rates in detection monitoring programs: false positive rates are computed per facility per annum, whereas false negative rates are computed per well per evaluation. There is no problem with the per facility/per well asymmetry: that is justified based on the consequences of making a wrong decision. I see no justification for the per annum/per well asymmetry. Indeed, the proper time frames are clearly established by two factors. First, the facility wide false positive rate *for the lifetime of the permit* is the most relevant indication of the facility's risk, not the rate per annum. Second, the false negative rate should be evaluated over a period of time during which it would be acceptable for contamination to be appearing at compliance wells without the facility actually taking some action. This is a complex, site-specific interval. In some cases it will be as short as the time elapsed between two evaluations. In many cases, though, it can safely span many years. By ignoring this distinction, the UG is effectively treating potential contamination at all facilities as

equally risky or harmful. That flies in the face of all other elements of EPA's risk-based approaches to environmental management.

As a result, I recommend that the UG *not* try to suggest that it is providing a comprehensive, universally applicable framework for applying statistical methods to RCRA groundwater monitoring. It should explicitly allow for creative application of new methods, provided they can be shown to meet RCRA performance standards and general EPA criteria. It could go so far as to point out (as it does in one or two places) other statistical methods that might be applicable, but which it does not comprehensively discuss or evaluate. Without such explicit acknowledgment, I fear that many regulatory agencies will simply not approve any proposal that does not fit exactly within the UG's framework.

b. Is the guidance presented in a manner that will be accessible to groundwater professionals with a limited background in statistics? Please explain your answers and offer suggestions, as appropriate.

I am impressed with the effort that has gone into getting the UG into its present form. It is remarkably free of errors of punctuation, grammar, and spelling. It contains many extensive passages that discuss important issues of groundwater monitoring statistics with clarity, completeness, and insight.

Nevertheless, it is equally clear that the UG is the product of many separate contributions and editorial revisions. The evidence lies in abrupt changes in writing quality, inconsistencies in format, and inconsistent usage of terminology. As a result, despite its many fine qualities, the UG could benefit from a systematic editorial revision aimed at making the writing style, the terminology, and formatting consistent throughout. Such a revision would help make this document more accessible to all readers, especially those not intimately familiar with statistical terminology and thought. For them, the writing in many places may still appear to be vague or confusing.

The figures in the UG will be very helpful once they, too, are thoroughly revised. They are not in as good shape as the text: almost universally they lack the information needed to be reliably interpreted by most readers. Most of them need descriptive captions, labeled axes (where appropriate), and other informative material. In my detailed comments later in this report I usually have not specifically mentioned the figures individually because these problems are pervasive: *without exception*, every figure in this document needs attention. To clarify this point, I will discuss a few figures below (see "Improving the figures").

Improving the writing

From experience with the previous guidance, I give great weight to clarity of writing and graphics because this avoids ambiguity. Ill-trained or misguided consultants can seize on ambiguous sections of guidance, even just solitary words or short phrases, in a search to find something advantageous to their client. Regulators themselves sometimes

misunderstand their own guidance due to misinterpretation of the same. Ambiguity does not serve the law or the environment well.

I comment at length in this report about elements of the text and tables that could lead to ambiguity, confusion, or misinterpretation, even though these elements may appear to be correctly structured and readily understood by the statistical professional.

Let us consider Section I. This is a key part of the UG because it establishes terminology and sets forth EPA philosophy and goals. It establishes a framework for the rest of the guidance. Unfortunately for the uninitiated, this section appears to use many statistical terms and concepts in multiple, distinct ways. I suspect that the careful reader, especially one not statistically trained, is likely to become confused.

The problem lies in getting the details right: defining terminology, using it consistently, using common English words in their conventional senses, making distinctions where they are important, and not implying false distinctions where they do not exist. Here are some examples from Section I that are characteristic of problems that recur throughout the entire UG.

- The UG states "the facility is required to conduct an assessment program identifying concentrations of hazardous waste constituents…" [at page 2-2, top]. One *identifies* the constituents themselves—presumably, most of them have already been identified, but new ones might also be through measurement of Appendix IX (or II) constituents. However, one *measures* concentrations and *estimates* their true values in the groundwater. Use words, especially verbs, appropriately.

- A *limit of detection* and a *practical quantitation limit* are different things. (The regulations themselves appear to confuse the two). Although later the UG mentions there is a distinction [at 10-1], here it quotes these regulations without comment [at p. 2-11, number 5]. Define key terms.

- Often, by using the passive voice, the UG inadvertently introduces ambiguity. This is especially the case when discussing recommendations. Here are some examples.

    - "To sidestep these conflicting sampling goals, at least two alternative strategies have been proposed" [at 3-21, top]. "… a second strategy has been proposed:…" [at 3-21, bottom]. Proposed by whom? The EPA? By the "RCRA statistical program" [at 3-22, bottom]? Does the UG recommend either strategy?

    - "While EPA believes that interwell tests still have an important role in groundwater monitoring, an alternative strategy that is often appropriate and recommended is *intrawell* testing" [p. 3-26]. Does the EPA recommend this strategy? Does the UG recommend it? Or is

the recommendation made by someone else (such as one of the "proposers" referred to in the previous quotations)?

Therefore, whenever possible, write statements in the active voice. Following this one recommendation alone will make much of the text far more readable.

- "Background" is used in multiple distinct senses. The UG originally calls it "a set of *baseline* measurements" [at 3-2], without specifically defining or describing "baseline." Is this a temporal background or a spatial background? Is it unit-specific, site-specific, or generic? Must it be fixed for all time (at least during the lifetime of a RCRA permit) or can it evolve over time? Some wells are referred to as "background wells" [at 3-7] without further clarification. In the sentence that follows the phrase "separate background or upgradient wells" could be construed either to (a) intend "upgradient" as a synonym for "background" or (b) make a distinction between "upgradient" and "background." (The same confusing distinction between "upgradient or background well screens" is made at [3-24].) Later [at 3-15 and 3-27], "background" is modified to "natural background" as if to make yet another distinction. Further on [at 3-18], "background" is used in the context of a discussion of *intrawell* tests, strongly implying that "background" data could be collected at downgradient or compliance wells. At [3-22], the UG refers to "background aquifers." How does this concept relate to the earlier uses of "background?" Are some aquifers "background" and others not? A quotation from the RCRA regulations [at 3-23, top] makes it clear that that these regulations make a distinction between "background" and "upgradient." The UG first explicitly acknowledges different kinds of "background" late in chapter 3 with the use of "intrawell background" [at 3-26] and "well-specific background" [at 3-27, top]. Clarifying the meaning of "background," and being consistently accurate with this terminology, is fundamentally important. Previous guidance and regulations suffered by not being specific about what constituted "background" in particular intended applications. The UG can improve by rectifying this omission. It might help to adopt clear, specific language to distinguish different concepts. Among others, the words "upgradient," "reference," "historical," and "natural" are available to help with that—provided they are used consistently and in a well-defined way.

- Equating "sentinel wells" with "compliance wells" (at 3-2) is ambiguous in a different sense. *Per se*, within the context of the UG only, it creates no ambiguity. However, I believe many readers will not agree with this equivalence. A sentinel well typically serves a purpose distinct from RCRA compliance, even though it might also be a compliance well. As another example of such a potential ambiguity, the UG uses the word "violation" [at 3-4 and 3-5] without definition. Is this intended to be a synonym for "statistically significant increase," or does it mean something else? Normally,

many groundwater professionals would take "violation" to mean some sort of RCRA permit violation, which could include more than an SSI; many would also not consider an SSI in itself to be a violation. A third example, slightly different from the first two, is the reference to "stratified-aquifer well" (at 2-18, top). One can only guess what this phrase might mean; I am not aware that it has appeared anywhere (it's not a common term). Here the potential ambiguity (if I am correct) derives from an unusual use of hydrological terminology. I therefore recommend that groundwater professionals, not specially trained in statistics, be asked to review the UG concerning its use of such "terms of art" to assure that they conform with commonly accepted usage and do not create potential confusion.

- The UG uses the technical word "distribution" [as "underlying distribution" at 3-11, "actual distribution" at 3-13, and "background distribution" later in Chapter 3] without definition. Although only infrequently used, this word appears in important contexts. It deserves a definition and discussion. I also recommend using modifiers (like "underlying" and "actual") in a consistent way, lest the untrained reader be concerned that important distinctions are being made. Describe and explain all statistical terms, no matter how elementary they may seem to the professional statistician.

The other chapters of the UG exhibit similar instances of potentially ambiguous writing and use of terminology. Rather than collect all such examples here, I have noted some of them in the chapter-specific comments below.

Improving the figures

To have any confidence that a figure will be correctly interpreted, it needs to stand alone: a reader turning to a figure for the first time must be able to understand it without having to read all the surrounding text.

As an example, let's apply this principle to the first figure that appears in the UG, Figure 3-1 [at 3-6]. It is not at all evident what this figure is showing or what mechanism it uses to do that. The figure indicates that time is on the horizontal axis, increasing left to right, but what is on the vertical axis? What do the "I-beam"-like graphics mean? Exactly how do they depict confidence intervals? Why are only three such graphics shown around the "increasing trend" line, which ought to have confidence bands? Do any of the times shown depict the future, or are they all times in the past?

To answer these questions, add appropriate labels to the figure and provide an explanatory caption. In this case, show explicitly that the vertical axes represent concentration, with higher concentrations to the top. In the caption explain that the "I-beam" graphics represent *two-sided* confidence intervals[1], with the bar spanning the

[1] Better yet, show one-sided confidence limits, because those are what the UG later recommends. The use of two-sided intervals will be quite limited.

entire range of the interval. Explain that a sequence of intervals is calculated, one interval at each "evaluation period" required by the monitoring program. Then describe the intended interpretation: in both panels of the figure, the first two intervals do not provide evidence that the facility is out of compliance, because their lower endpoints are not above the GWPS, whereas the last interval provides statistically significant evidence that the facility is out of compliance, because its lower endpoint is above the GWPS. Finally, reference chapters 15 through 17 where these matters are discussed in detail. This is the additional material that is needed to turn this rough sketch into an informative illustration.

Consider, as another example, Figure 4-1 [at 4-4]. Any statistician will immediately understand and interpret this correctly, but I suspect many readers will not. It needs a caption that explains how *areas* beneath the curve represent probabilities; how to interpret the shaded area; how many "sampled values" are involved; what the numerical range of those sampled values was; and what a typical value for benzene is. Either explain the mysterious negative concentrations shown at the left, or leave them off the figure altogether. Change the misleading "probability" label on the vertical axis: it is really probability per unit concentration, not probability itself.

In general, for every figure in the UG:

- Provide an explanatory caption that enables the figure-plus-caption to stand on its own.
- Clearly label any axes, providing the name of the variable it shows, some indication of sign and magnitude, and units of measurement.
- In the caption explicitly describe how the symbols (that is, geometric objects and the graphical methods used to draw them and differentiate them from each other) represent information.
- In the caption, explicitly describe the intended interpretation. Explain exactly how the figure illustrates the point that is being made.

In short, leave nothing to chance: make sure all readers will correctly interpret every figure.

Improving the tables

Every table needs additional material to assure its proper use. This material should include:

- Explicit statement of the table's purpose.
- A clear description of the variables.

- Cross-references to the section(s) of the UG where use of the table is described.

- One or more brief worked examples showing the use of the table.

- Where appropriate, an indication of how interpolation (and, where possible, extrapolation) should be performed, with a worked example.

- Description of the method used to compute the tabulated values, along with an indication of whether they are exact or approximate.

- Warnings about any necessary statistical assumptions (such as Normality) or limitations.

- References to published sources, including the sources of these tables, descriptions of the algorithms, and tables that have greater scope or precision.

In addition, avoid abbreviations in the titles and descriptions: it's worth taking an extra line of text in order to be clear and unambiguous.

To see why this material is important, go (for example) to Table 6-7 [at C-10]. How many people would have any idea what to do with these numbers? The principle to follow, similar to that for figures, is to make the tables stand alone, so that they can be used by any occasional, non-expert reader with minimal effort and maximal reliability and accuracy.

2. Overall, is the document well organized and cross-referenced in a manner that will help users apply the guidance? Please explain and offer suggestions, as appropriate.

It is evident that much thought and care have gone into the organization and cross-referencing of this document. It's in good shape.

Below, I make some particular suggestions for improvements, but first I want to bring up an important issue. It seems that the UG is trying to address two interrelated but greatly different needs, which we might characterize as *permit writing* and *test execution*, or more generally, strategy and tactics. During the permit writing phase the reader needs to develop a groundwater monitoring program that integrates sample collection and analysis with statistical testing and decision making. It is at this stage that any historical and background data will be extensively tested and characterized. Here is where considerations of power and significance are paramount. Once the permit is in place, data are collected, analyzed, and evaluated according to the permit. Power and significance no longer need to be evaluated (at least not routinely). At this stage the reader will likely use the UG primarily for two things: making sure the statistical test is correctly carried out and checking whether the data suggest that statistical assumptions made during the permit writing process may have been violated.

It therefore is logical that the UG should reflect these different needs within its organization. However, I see no recognition at all of this distinction within the present document. Consequently, the reader of the UG in its present form does not get useful guidance concerning exactly when to conduct tests of distribution, autocorrelation, etc., and when not to. Guidance on key monitoring design issues, such as that one can use different statistical approaches for different monitoring constituents, is embedded within the various (and quite extensive) chapters of Sections II, III, and IV. As such, it seems inevitable that most readers tasked with designing a monitoring program will either feel compelled to read the entire UG (at 570+ pages, not counting appendices, tables, references, or prefatory material) or—more likely—will just give up due to the daunting nature of the task and not receive the benefit of the really impressive expertise that has been incorporated.

These considerations suggest a reorganization of the UG based along practical lines. Where should a reader go to learn how to select an appropriate set of statistical tests? That should be one section. It would be comprised of much of the current Section I along with much of the background and prefatory material appearing in chapters 6 through 17, as well as choice elements that are buried in detailed procedures and examples. Where should a reader go for guidance on writing the permit language itself? That section needs to be written. Where should the reader go for the algorithms to execute the tests? That section would comprise the procedures and examples of chapters 6 through 17. Where should a reader go to learn how the UG differs from the previous guidance? That should be a separate section, comprised of the "historical notes" presently scattered throughout many individual chapters. How about learning some of the theoretical motivation behind the new guidance? Put that into an appendix.

This would include the results of Monte-Carlo simulations that appear in separate chapters.

This reorganization could make the UG much more approachable. A section that collected all algorithms (procedures), examples, and case studies would be about the same length as the 1989 guidance: mercifully, much shorter and more approachable. A section discussing the process of monitoring system design, test selection, and the EPA's interpretation of the regulations would contain the most readable parts of the UG (in my opinion) and would be about the same length. By itself, it would make a wonderful introduction to and overview of groundwater monitoring statistics and the EPA's implementation of the RCRA performance criteria.

I anticipate that anybody who has been involved in developing the draft UG would be reluctant, at this relatively late date, to undertake such a reorganization. Without doing this, though, I think the sheer bulk and complexity of this guidance will work against its widespread use, especially among facility operators and their consultants. Note that I am suggesting re-organization, but not extensive rewriting. Indeed, there is enough redundancy built into the current structure that this re-organization would probably eliminate about ten percent of the pages.

The following comments on organization are relatively minor in nature.

In addition to Appendix E, an index or table of symbols and abbreviations would be welcome, especially to those who wish to consult only portions of the guidance infrequently.

It would help for the Table of Contents to show the major section divisions, because the Executive Summary refers explicitly to these divisions.

Many examples and procedures refer to other ones appearing in other chapters. It would help if they systematically contained cross-references by page number.

Many procedures assume the reader understands exactly what formula should be used for a mean, a standard deviation, a log mean, a log standard deviation, a coefficient of variance, and even a log coefficient of variance. Because these occur so frequently, consider presenting their formulas in a single table and cross-referencing that table wherever possible.

There are minor but important inconsistencies from one chapter to another. It would help, for instance, to present all example data in the same way. Many examples contain sample dates, but they appear in a myriad of formats, many of which are not even immediately recognizable as dates. Adopt a uniform convention for presenting example data so that unnecessary variation does not cause confusion.

In many cases there is no clean division between the text and the formal description of a "procedure" or "example." Quite a few procedures and examples introduce new ideas, such as recommended values of alpha, that really belong in the background discussion.

Some sections describe essential parts of procedures, including formulas, within the text; those formulas should appear in the procedures. (I point out many instances of such lapses in the detailed chapter-by-chapter comments below.) In short, for the UG to serve as a good reference manual, *all procedure descriptions should be completely self-contained*: the reader should be able to go directly to the first step and be able to follow it successfully through to the last step without hunting around in the document for necessary formulas and definitions.

3.I. For Section I of the Unified Guidance, please address the following questions :

a. Does Section I meet the stated objectives described in the Introduction to the Charge, above? Please explain.

This section meets its stated objectives. It is logically and clearly organized. It covers an appropriate range of topics.

It touches on some of the most important issues that have come up and will continue to come up, including:

- Why the EPA regulates a minimum false positive rate (4-17).
- The distinction between statistical power expressed in terms of standard deviations and in terms of concentrations (Section 3.5: this is very well done).
- The importance of understanding spatial variability (many places).
- The distinction between statistical significance and environmental significance (4-13, top).
- A formal statement of what the EPA considers to be an acceptable annual sitewide false positive error rate (3-11).
- A clear discussion of when intrawell testing should be considered as opposed to interwell testing (Section 3.7).
- The flexibility available to craft permits that allow the monitoring program to improve over time in a structured way (3-32, bottom) by (among other things) "temporarily defer[ring] comparisons" and updating the background sample set.

b. Does Section I cover an appropriate range of topics? Are there any key topics that are missing or that should be emphasized or described in further detail? Please explain.

There are some more key topics that would be worth discussing in this introductory section. They include:

- The possibility of correlated variation, how to recognize it, how it can affect the tests, and how to cope with it. Although section 9-2 discusses correlation in more detail, it might be well to alert the reader about its importance here in Section I of the UG. Section 3.4.1 does mention correlation, but does so in passing and without making distinctions (such as temporal, spatial, and other): I am instead advocating a more generic discussion of how correlation

can change the actual Type I and Type II error rates of the tests and what in general one might do to identify its presence.

- The need to graph the data. Regardless of what test is chosen, there is no substitute for effective, routine graphing of monitoring data as they are collected. Always graphing the data should be a strong EPA recommendation, to be repeated throughout the UG. (If this guidance could do just one thing to improve decision making at RCRA facilities, it would be to require, in the strongest possible way, routine submittal of clear graphical displays of monitoring data. Most facilities do only what they are required to do, and so they wind up using a series of mindlessly executed statistical tests as the sole basis of all decision making in their RCRA monitoring programs. Make them *look* at the data!)

- Time frames. This is especially important in section 3-5. All discussion of statistical power appears confined to the power of a single test to identify an SSI, whereas the sitewide false positive rate is measured during a fixed period of one year. Why the asymmetry? If groundwater conditions have changed, but the testing during the next monitoring round fails to detect that, then (assuming conditions do not get better), the testing in subsequent rounds increases the probability of a detection. A proper discussion of power and false negative rates needs to account for the time it takes to detect an increase, not just the probability of detection during one round. If the EPA is going to start making recommendations based on effect size—and I believe that's a very good idea—it should go all the way and acknowledge that some situations require rapid detection and response, while at other sites relatively slow detection (perhaps over the course of three or four rounds) would be more than adequate. (The 1989 guidance made a start at this in its analysis of in-control and out-of-control rates in a control charting context.)

- I would like to quibble with a distinction made at several places, including at pages 2-19 and 2-20, concerning "physical" and "statistical" independence of samples. Wherever this distinction is made, there seems to be an implicit assumption that contaminants flow without dispersion or diffusion. This of course is never the case. Accordingly, it is difficult to establish exactly what might be meant by "physically different portions of an aquifer" (as at 2-20). This over-simplistic view of contaminant transport could lead people to conclude that "independent" samples could be collected over much shorter intervals than in fact they can be. It could be helpful to include an informed discussion of this issue somewhere in Section I. It's nice to see that later sections do exhibit a better understanding of this phenomenon, to the point of acknowledging that any sampling more frequently than quarterly would be unusual.

In Section 3.4.2, one might add several factors to the list, including the possible presence of pumping (or other groundwater extraction) both onsite and offsite, any change in well

purging and sampling procedures, and the possibility of preferential flow in the aquifer (secondary permeability). In Section 3.4.3, another factor to consider is whether changes in groundwater quality (not due to a release from the regulated unit) might engender changes in seemingly unrelated constituents. A classic example—and one likely to be encountered frequently beginning in 2006 due to the lowering of the arsenic standard—is the mobilization of naturally occurring metals, such as arsenic, from the rock or soil matrix through changes in pH and other (benign) groundwater quality factors.

c. Is the material in this section organized and presented in a clear and concise manner? Please explain.

Please see my general response to Question 2 above.

3.II. For Section II of the Unified Guidance, please address the following questions :

a. Does this section meet the stated objectives described in the Introduction to the Charge, above? Please explain.

It systematically covers the statistical assumptions listed in the RCRA performance requirements, one chapter per assumption: statistical distribution, homoscedasticity, correlation, seasonality, below-PQL values, and even outliers. The table of contents alone indicates this section is likely to meet the objectives.

Chapter 6 contains a detailed, thoughtful, and accurate discussion on the topic of default distribution assumptions for groundwater monitoring. The criticism of previously-recommended tests, such as the Chi-square test [at 6-18], is very well done. Chapter 10 is similarly written in a thoughtful, helpful way.

b. Does Section II cover an appropriate range of topics? Are there any key topics that are missing or that should be emphasized or described in further detail? Please explain.

General comments

This section is silent about two key topics: how often to perform diagnostic testing and which data to use for this purpose. For instance, at [6-10, top], is the UG recommending *routine* diagnostic testing (with every sample event), diagnostic testing at planned intervals (such as at two-year evaluations), or only during test selection (during the permit development process)? When testing is performed, is it to be performed only on background data or on all data?

If routine diagnostic testing is intended, then P-values and power estimates for *every* test in the guidance would need to be recomputed. Take, for instance, the simple situation in which a permit proposes to follow a two-phase approach of first testing for normality *vs.* lognormality and then computing a prediction limit of size α. This composite procedure almost surely will not have the intended size, nor will it have the power computed for either the normal or the lognormal prediction limits. Therefore I would hope that extensive diagnostic testing would be reserved for the test selection process (during the permit development phase) and that only a minimal battery of diagnostic tests, such as tests for extreme outliers, be recommended for routine, ongoing use. If the EPA agrees with this approach, then the UG must clarify the times and frequencies with which the various diagnostic tests ought to be applied and also it should supply the recommended responses to the diagnostic results. For example, if the UG recommends testing for approximate equality of variances every time an ANOVA is applied, then it should also state what exactly should be done when the test suggests variances are unequal.

Chapter 6

This chapter contains a thoughtful and thorough discussion of the role of the Normal distribution in groundwater monitoring tests [at 6-1 through 6-11]. This is very well done.

Chapter 7

Chapter 7 presents good, effective methods of testing for equality of variance. It does not cover all the ground it should, though. Tests of homoscedasticity are also needed for time series data such as monitoring values and residuals relative to trend lines. The box plots and Levene's Test are readily adapted for this purpose by dividing a time series into groups (such as halves or thirds) and applying these diagnostic tests to the groups. However, the UG, although it recognizes that such tests are needed[2], does not describe such applications. I recommend that it do so.

Chapter 8

This chapter presents two formal tests for outliers: Dixon's test for a single outlier in small data sets and Rosner's test for up to five simultaneous outliers (high or low) in larger datasets. However, both tests use the Normal distribution as a reference, so that Normality or the ability to transform data to approximate Normality (apart from the outliers) is necessary. This chapter could therefore be usefully supplemented by a robust, nonparametric (if less formal) test of outliers. I recommend Tukey's fences (John Tukey, *Exploratory Data Analysis*, 1977). These fit well with chapter 7 of the UG, which recommends and describes box plots. Tukey's fences can be derived from the box plot calculations. They are easy to calculate. One can even estimate them accurately by looking at a box plot. They can identify an arbitrary number of outlying data. They do not require the user to specify the number of outlier in advance. They identify and classify outliers into four groups: "near" and "far" outliers, high and low. This classification can be used to guide subsequent actions.

Chapter 9

Readers attempting to apply the material in this chapter might often be frustrated, as its own examples illustrate. For instance, the ranges of standard deviations for *both* the raw and logged data in Example 9-1 [plotted at 9-13 and 9-14, respectively] are so great that they fall into the "severe drop in power" regime noted by Milliken & Johnson [at 9-16]. What is the poor reader supposed to do in this case? Indeed, my calculations indicate that *no* power transformation will stabilize these variances sufficiently.

More generally, when all diagnostic tests fail—data and residuals do not appear to be Normal or Lognormal; variances, despite data transformations, are not even

[2] *E.g.*, "Equality of variance is assumed, for instance, when using prediction limits in … intrawell comparisons. … [I]t is assumed that the well variance is stable over time when comparing intrawell background versus more recent measurements." [At 7-1, top.]

approximately stabilized; trends are nonlinear; nondetects are numerous and have wildly varying detection limits; and so on—what is one supposed to do? Often, the UG is silent on these points. In places it suggests consulting a professional statistician and in others it suggests completely changing the statistical test, but it does not do this everywhere. Maybe it would help (perhaps in Section I) to provide some default recommendations for steps to take when nothing seems to work. Even though this Guidance is not intended for statisticians, it still could be useful to point to additional approaches that statisticians could validly consider: other parametric distribution families, fitting nonlinear trends, robust regression, imputation of NDs with varying censoring limits, and so on. (Just "for reference," as the UG kindly points out [at 9-28].)

Chapter 10

This chapter needs a section on how to treat nondetects when conducting tests for trend and control charts [Chapter 14].

c. Is the material in Section II organized and presented in a clear and concise manner? Please explain.

Chapter 6

Some of the material in Chapter 6 is technical. Its role is to support an explanation of changes. It would help to put this into an appendix. This includes the material from 6-6 through 6-15 reporting Monte-Carlo analyses of prediction limit calculations.

The presentation of probability plots in Section 6.4 [at 6-20 *et seq.*] seems unable to decide whether to describe only Normal probability plots or probability plots in general. Only small changes are needed to make this presentation quite general (remove references to the Normal distribution). I recommend making such changes so that the UG explicitly opens the possibility of using probability plotting for distributions other than the Normal or transformed-Normal.

A rationale for the recommended test sizes in Step 6 of Procedure 6.5.1.2 [at 6-24] would be welcome. I understand the need for decreasing levels of α and am not quarreling with that, but the particular schedule advocated here is a bit of a straitjacket. Readers ought to be told *why* the significance level should be chosen to depend on *n* and provided *guidance* in making an appropriate choice, rather than given an unjustified, inflexible prescription for this choice. Such a discussion does not belong here, buried in a step within a detailed procedure. It is applicable to all tests of distribution. It belongs in the Summary section (6.1) or in the subsequent discussion (Section 6.2).

Please remove the word "unfortunately" from the sentence [at 6-33, top] or rephrase the introductory clause so that the word "unfortunately" is not applied to the recommendation in the UG!

Replace every occurrence of "Wilk" (there are many) by "Wilks," which is the correct spelling of Samuel Stanley Wilks' last name.

Chapter 7

The language in Chapter 7 sometimes becomes too restrictive. The UG envisions testing equality of variances of groups of well data over time. Thus, the discussion of Levene's Test needs to consistently refer to "groups" rather than the more restrictive "wells." This change needs to be made in three places [at 7-6 and 7-7].

There is a tendency for important guidance to be buried within procedures or examples[3]. Procedure 7.3.2 [at 7-7] provides one instance, where the description of the procedure at Step 8 is interrupted to provide guidance on the appropriate significance level to use. This material would fit better in the introduction to Levene's Test [near 7-5, middle]. Review the entire UG for places where important, general guidance first appears within procedures or examples and then move that guidance, with more extended explanation, into appropriate introductory or summary sections.

Using "1.5-2 times" [at 7-2 in two places] is unnecessarily vague. Why not just say "twice?"

In Section 7.2 [at 7-2 and 7-3], it might help readers to point out that at least three distinct and slightly different methods of constructing box plots are commonplace. The procedure for computing the quartiles [at 7-2 and 7-3] is relatively uncommon and is likely to produce plots that differ slightly from those produced by software or by following textbook procedures. Software is most likely to linearly interpolate the quartiles, rather than halving the ranks. The original method (Tukey's, *op. cit.*) uses "hinges" instead of quartiles and some software still uses this technique. (The hinges are the order statistics at depths $(\lfloor (n+1)/2 \rfloor + 1)/2$, where $\lfloor\ \rfloor$ designates the floor function). It would be nice to see the UG acknowledge that the differences introduced by these various conventions are acceptable.

By the way, it's fairly rare—and an indication of crudely written software—for the whiskers in a box-and-whisker plot to extend to "extreme" values [at 7-4, bottom]. They usually extend to the most extreme of the *non-outlying values*, reserving discrete symbols to depict any outliers.

Chapter 8

Some key guidance is embedded within the steps of the procedure descriptions [Step 5 at 8-9 to 8-10 and Step 6 at 8-12 to 8-13]. This concerns actions to take upon identifying outliers. Move this guidance to the "Summary" or "Basic Strategy" sections and eliminate the repetition.

[3] The first was the recommendation concerning significance levels to use in tests of distribution, located in Step 6 of Procedure 6.5.1.2. Subsequent chapters exhibit many more instances of this unfortunate technique.

In section 8.4.3 [at 8-10], the UG suggests "it may be helpful to attain another sample in order to verify or confirm the initial measurement." Good advice, but what should one do if the new sample does confirm the initial measurement? Should both results enter the dataset and be used for analysis? Should their average be used? Should the original be replaced by the confirmation result? Should the confirmation result just be ignored in that case? The UG needs to address these questions. Consider writing a separate chapter providing guidance concerning maintaining a database of groundwater monitoring results. This has always been an important issue, but finally—after over twenty years—most facilities are aware of it and want to do something about it.

Other guidance is provided quite casually, in passing, at the beginning of Example 8-2; namely, to conduct Dixon's test at a 0.05 level of significance. Instead, present a strategy for choosing significance levels in the "Basic Strategy" section. Provide guidance for adjusting this level of significance when evaluating many independent batches of data. Do not overlook the fact that informal evaluation occurs with *every* batch of data, whether or not a formal test is performed, so that the adjustment should account even for the informal evaluations that are performed. (For this reason, I think it is rare that a significance level of 0.05 will be appropriate.)

Chapter 9

The presentation of this chapter is not as clear or accurate as the chapters that precede it.

- In some places the verb choice is misleading: for example, "identifying" instead of "assessing" [at 9-1, middle]; "delineate" instead of "differentiate" [at 9-10, middle]. Change these to make the text more accurate.

- The reader is assumed to understand terminology that has not yet been used or defined in the UG. For example, Example 9-3 [at 9-22] proposes "to compute adjusted intrawell prediction limits" without defining these or referring to a later section in the UG that describes them; observations (that is, numerical concentrations) are (mysteriously) described as having "position and magnitude" [at 9-27]; the "efficiency" of a statistical test is discussed [at 9-27, bottom], and "effective sample size" is mentioned [at 9-33, bottom], without definition or any indication of its potential use. Either avoid such terminology, provide cross-references, or explain it wherever it appears.

- In most places, all correlation is assumed to be *positive* correlation. The possibility of negative correlation is almost completely neglected. At times the phrase "positive correlation" appears [as at 9-5 and 9-8], indicating that the UG is aware of negative correlation, but the recommended test does not accommodate this: it only tests for significant positive correlation. Include negative correlation in the discussion. Provide an example. Modify the rank Von Neumann test to detect significant negative correlation. (Negative correlation does occur in groundwater monitoring data series, most often through twice-yearly sampling of seasonal parameters.)

- The UG provides confusing and contradictory recommendations. In the discussion of seasonal effects, for instance, we are told authoritatively and without reservation that "both the upgradient and downgradient data should first be de-seasonalized prior to statistical analysis" [at 9-39]. One paragraph later, however, the UG asserts that "[c]orrections for seasonality should be used with great caution…" [at 9-39, bottom]. Overall, the UG is properly cautious and provides many caveats concerning the use of seasonal correction, but I am concerned that someone could seize on a clear statement (like the former) to the exclusion of all other. Since correcting for seasonality is a drastic step to take, has a high potential for error, and is rarely needed in practice, I would be happiest to see the entire section (9.4.5) reduced to a single recommendation: "If seasonality is suspected … the user should seek the help of a professional statistician" [at 9-40, top]. (Other references to seasonality in later chapters should also be updated to reflect any changes to the guidance in Chapter 9.)

- The discussion leading up to the rank Von Neumann procedure [at 9-27] lacks the usual clarity and accuracy apparent in earlier discussions in the UG. Statements such as "each sample measurement is utilized more than once in the computation of any autocorrelation," "the first-order autocorrelation in *dependent* data will tend to be positive," and "the rank von Neumann ratio test statistic is built around the sum of differences between the *ranks*," are either questionable or demonstrably wrong. This entire section would benefit from heavy editorial intervention, rewriting it for clarity and accuracy.

- Example 9-6 [at 9-35] is incomplete. It should display a standard ANOVA table. Because it intends to compute a prediction limit (PL), it should show how that PL would be computed based on the estimates it has derived.

- The procedure in Section 9.4.3 needs much more guidance concerning which significance levels to use. This is a complex procedure involving a sequence of several tests, so choice of levels is subtle and tricky. It asks the reader to test normality of residuals, to test normality of group means, to test equality of variance, and to test for significance of the F-statistic. It does contain some guidance, presented in passing during the step-by-step description of the procedure [at 9-33]. Remove this guidance from the procedure description, include it in the preceding material, and provide additional guidance to help the reader select an appropriate collection of significance levels for the individual tests.

- The intended applications of the guidance in Chapter 9 are not at all clear. In many cases it appears that only certain kinds of prediction limits are being considered: intrawell PLs [at 9-21, top] and interwell PLs [at 9-35]. Please clarify the circumstances in which the tests in this chapter should be considered.

- I am concerned that in this chapter, as in some places elsewhere, the UG is using test size as a surrogate for effect size. A clear example appears in the recommendation to conduct the rank Von Neumann ratio test at a size of $\alpha = 0.01$, "since only substantial non-independence is likely to degrade the results of subsequent statistical testing" [at 9-29 top]. The amount of non-independence (perhaps expressed as a ratio of the true long-term variance to apparent short-term variance) surely can affect follow-on tests, but that amount is not the same as the significance level of the test result. Consequently it appears that the UG is unnecessarily limiting readers' options due to this conceptual fallacy. The UG should provide broader ranges of recommended test sizes, especially for diagnostic tests, to allow for appropriate application as circumstances warrant.

- This chapter contains a perceptive "note on correcting for linear trends" [at 9-42 and 9-43]. In support of this note's conclusions, it might help to acknowledge an important, commonplace phenomenon. In most groundwater monitoring systems where multiple upgradient wells exist, then (due to the fact that the compliance wells almost always outnumber the upgradient wells) it is often the case that some downgradient wells will exhibit smaller mean concentrations than any upgradient well. There are many reasons why that might occur, both for natural and anthropogenic ("synthetic") constituents. A concern among facility operators in this circumstance is that possible long-term secular changes in groundwater flow or quality could cause concentrations in a downgradient well slowly to increase over time. For instance, a slight change in groundwater flow paths could cause groundwater passing through a high-concentration upgradient well to start traveling towards a lower-concentration downgradient well (which previously had not intercepted that water). Whatever the true reason may be, such trends in downgradient data are frequently observed. They exhibit themselves as the presence of a downgradient well exhibiting significant upward trends in concentration, but the concentrations always remain below those measured in immediately upgradient wells. Even when such a trend is significant, it is not valid to infer that the regulated unit (lying between the upgradient and downgradient well) is causing the increase. Moreover, such a situation appears unlikely to represent an environmental threat, because the downgradient concentrations remain lower than the upgradient concentrations, which presumably are acceptable. The possibility of such a scenario demonstrates why adopting only intrawell tests, without also including some (interwell) comparisons to background, can be foolhardy.

Chapter 10

Organization

Parts of this chapter, especially parts of the "Overall Framework" [at 10-1 through 10-14] are clearly presented and informative. The subject matter, though, creates problems for organization, more so than in chapters six through nine. There are many aspects of this topic to cover: definitions, past guidance, new findings, and applications to different forms of tests and analyses (ANOVA, prediction limits, identifying trends, etc.). The subject is further complicated by the variety of censoring that can occur, whether via nondetects, below-PQL data, or via laboratory reporting limits; by the proportion of censoring; by the patterns of censoring among wells and over time; and by variations in censoring levels (multiple detection limits).

The UG exists primarily to solve two basic problems that users will routinely confront: (1) selecting an appropriate battery of tests to meet the RCRA regulations and (2) carrying out those tests as data are collected. Much of the material in this chapter appears appropriate for the first use, test selection. Where the material is specific to a particular test, as in "Non-detects in T-tests and ANOVA" [at 10-10 *et seq.*], it might serve the user better by appearing in the part of Section III devoted to that test. Otherwise, with the current organization, a user would first go to the test-specific material in Section III, and then would have to hunt through Chapter 10 to learn how to handle nondetects (and, similarly, hunt through Chapter 7 to find an appropriate method to test equality of variance, etc.) Therefore, consider moving test-specific guidance out of Chapter 10 and placing it in the appropriate part of Section III. Thus, Chapter 10 could describe Cohen's and Aitchison's methods in general, leaving their specific application (such as to computing PLs) to later sections. This would generalize the scope of the chapter and help avoid confusion.

The Simulation Results

Move the description of the simulation into a technical appendix. This would make it possible to publish the complete results, rather than the "average" statistical power (which I suspect must be the arithmetic means of the powers observed over various choices of underlying distribution).

3.III. For Section III of the Unified Guidance, please address the following questions :

a. Does Section III meet the stated objectives described in the Introduction to the Charge, above? Please explain.

This section falls short of meeting the objectives in a spectacular way: it clearly allows a facility to perform as many *post-hoc* tests as it would like and to choose whichever results it likes. This is nowhere clearer than in Example 13-7 [at 13-63 through 13-65]. This example proceeds by showing how a facility could run *six different tests* at each compliance well and choose the tests that cause all the compliance wells to pass.

Evidently, the intent of Example 13-7 and of many (if not most) of the examples offered in Section III is to illustrate procedures to select tests that subsequently will be prescribed by the RCRA permit. However, *the UG presently contains no language that clearly differentiates the process of test selection from test execution*. I believe the entire UG could systematically and justifiably be mis-interpreted in this regard, especially by naïve or unscrupulous readers.

I am also concerned that Chapter 13 requires readers to perform calculations that are beyond the capabilities of most of the intended audience to perform: namely, the Monte-Carlo simulations of power. In principle these are not difficult to carry out, but they are very difficult to check for accuracy and to carry out correctly, except by highly qualified and experienced statisticians. Is it really the intention of the EPA to drive all facilities to using consultants to help them develop RCRA permits? If there exist common situations where the EPA would not require or recommend explicit power calculations, then please describe these clearly and simply, so that the intended readers can readily determine their ability to follow this guidance.

b. Does Section III cover an appropriate range of topics? Are there any key topics that are missing or that should be emphasized or described in further detail? Please explain.

Provide a reasonable analysis of sampling and analytical costs

Chapter 13 frequently acknowledges that resampling plans can create variable sampling and analytical costs, but its analysis is limited and practically useless in this regard. Because most facilities will not be detecting contamination every time, and because the default hypothesis during detection monitoring is that there is no contamination, the cost to perform the sampling and analyses when contamination is present is practically irrelevant. Instead, the reader will be keenly interested in *the (long-run) expected numbers of resamples and re-analyses* that will be needed with any formal resampling plan[4].

[4] It was a pleasant surprise to find that the nonparametric prediction limit software attempts to quantify the number of resamples needed for certain plans. This is a step in the right direction. However, for reasons about to be discussed, it is not sufficient.

Examples of this limited and frequently misleading analysis appear at [13-48, bottom], [13-50, bottom], [13-53, bottom], [13-54, top], [13-75], [13-76, top], [13-77], and [13-79, top].

I refer to "resamples" and "re-analyses" specifically because their expected numbers will differ from each other at most medium to large facilities. The main costs are associated with (a) the cost for a sampling team to visit a well, purge it, obtain one or more physical samples of water, and ship them to the laboratory; and (b) the cost for the laboratory to conduct a test (often producing a suite of analytical results) on one physical sample. A well will need to be revisited whenever *one or more* resamples is required of it. With a 1-of-2 resampling plan, for instance, if c independent *initial comparisons* (for c independent parameters) at a well have an expected positive rate of α^* under the null hypothesis, then the rate at which that well will need an initial resample is going to be $1 - (1-\alpha^*)^c$. The expected number of *re-analyses* of that resample is a slightly more complicated expression (depending on the number of analytical parameters in each test suite) that ranges between $1 - (1-\alpha^*)^c$ and $c\alpha^*$. These values have to be summed over all compliance wells. The evaluation of these expectations is feasible, although more complicated, for more complicated resampling plans. Nevertheless, *estimating expected costs of resampling and re-analysis* is a critical part of monitoring test selection and design for every facility.

Provide clearer, more explicit guidance about resampling procedures

The preceding considerations raise related issues. For instance, it is not at all apparent that the assertions about numbers of samples to collect for PLs of medians are correct. A median-of-three test is really a two-of-three test. Thus, a facility can frequently make a decision about a median (*i.e.,*, whether it exceeds a PL or not) by obtaining two samples, not three as prescribed [at 13-62, for instance]. If the smaller of the two results exceeds the PL, the comparison is positive; if the larger exceeds the PL, the comparison is negative. Only when the results straddle the PL is a third sample necessary.

Let's think this through in the case of a median-of-three plan with one resample. To be specific, let this plan call for twice-yearly evaluation based on samples obtained at regularly monthly intervals, so that up to six (presumably independent) samples can be obtained during each evaluation period. A PL is computed from historical background once and for all. Thus, the test can be conducted immediately upon receipt of the measurement of each sample, which typically will occur before the next sample is taken. This lets the facility *sample contingently based on the previous results obtained during the evaluation period.* Routinely, the first month's samples are collected and analyzed and then the second month's samples are collected and analyzed. The facility is now about to collect the third month's samples. Suppose that the first two measurements at a well exceed the PL. We conclude the median exceeds the PL. Thus, we already know that resampling will be needed. Several options now seem available: (i) collect the third month's sample (in its role as the third of the first group of three) and analyze it; (ii) collect the third month's sample but do not pay for an analysis, because it is already known that the median of the first three months' measurements exceeds the PL; (iii) collect the third month's sample, but count it as *the first of up to three samples to be obtained for the resampling*

process; and (iv) do not collect any sample during the third month, because (as in (i)) it is already known that the median of the first three months' measurements exceeds the PL.

Most facilities would opt either for (iii) or (iv), because the resampling cost in (i) and (ii) and the re-analysis cost in (i) are unnecessary for decision-making. From the point of view of obtaining timely decisions, (iii) would be optimal. Does the UG allow this? Whether it does or not, the UG must discuss the options explicitly. For instance, if a permit is written without specifying which option should be taken in every possible instance, consider what might happen when the facility intends to take option (iii) but then discovers that the third month's result also exceeds the PL. Would it not then be tempted to assign the third month's result to the *first* group of three—as if it had intended to take option (i)—so as to maximize the chances that the median of the next group of three will fall below the PL? This illustrates why every permit must be clear and explicit about procedures for resampling, analyzing re-samples, using their results in the evaluations, and incorporating the results in the ongoing monitoring database. It shows why the UG needs to provide clear and explicit guidance about all these aspects of data collection when resampling is a formal part of the statistical testing.

c. Is the material in Section III organized and presented in a clear and concise manner? Please explain.

Chapter 11

The rationalization appearing in the "Historical Note" [at 11-3 through 11-4] concerning appropriate circumstances for using a t-test appears to conflict with earlier statements in the UG. Whereas the UG had stated that the intended FWFPR is a maximum of 10% *per annum*, here the UG asserts a value of "10% probability per evaluation of any … false positive" is acceptable. In most cases this would create 19% to 34% FWFP rates. Please resolve this contradiction. (At the same time, much of the text in this chapter could benefit from a thorough editorial revision for clarity and conciseness. The redundancy in a phrase like "comparison registering as a false positive when there is no actual contamination" is palpable.)

The definition of the t-statistic [at 11.2] is in terms of symbols (the means, variances, and counts) that have not been formally defined. Please do so. In particular, show clearly that the variances are the unbiased estimators rather than the MLE estimators. This will help avoid confusion.

It will help some readers to indicate more clearly in Example 11-1 that "Log" refers to the natural logarithm, not the common logarithm [at 11-10, table heading].

The statement in Step 5 [at 11-11] is disturbing: "The fact that the conclusion [varies] based on a small change [in] the significance level should be troubling." Once a desired significance level is established, test results are either significant or not. A result that is just barely significant is still significant and one should not be troubled about that at all: one has to draw the line somewhere. At any rate, such a circumstance says nothing

about the suitability of the test (which is the point that the UG is attempting to make). The issue here, I think, is that the result is *sensitive to inconsequential changes in the data themselves*. For instance, an almost infinitesimal change in the oldest downgradient result from 0.5 to 0.3 changes the conclusion of the test. *That* is a valid point, one worth making in general. Indeed, it could be useful to provide general guidance on sensitivity analysis (perhaps in Section II).

The statement [at 11-12] "finding a larger geometric mean … in a downgradient well when compared to background also implies that the downgradient *arithmetic mean* is larger than the background *arithmetic mean*" (when both groups have a common population variance) is incorrect. It appears to be based on a confusion between sample variances and population variances. Even when the population variances are equal, it is unlikely the sample variances will be equal. Thus, it can happen that the relationships between estimated AMs and GMs will be reversed (*i.e.*, background GM > downgradient GM but background AM < downgradient AM, or *vice versa*). This will happen 20% of the time when background and downgradient data have identical Lognormal distributions with unit log SD, for example. Please review this section for accuracy and logical reasoning, making appropriate changes. In particular, please re-think the conclusion (at the bottom of 11-12), because it appears to be wrong.

The minor change in notation [at 11-13] is potentially confusing. Whereas before we had background, downgradient, and variables with "BG" and "DG" subscripts [at 11-5], now we have "first population," "second population," and variables with "x" and "y" as subscripts. There is no need for this or apparent advantage in doing it. Please maintain a consistent notation and terminology throughout this chapter.

Before presenting Example 11-2 [at 11-14], consider providing a formal "procedure" statement, to be consistent with the rest of the UG.

The discussion of the Wilcoxon test is concise and accurate. The UG needs to take a little care with the laboratory qualifiers, though: there are several conventions in common use. In particular, some "J" values are estimated not because of detection limit problems, but because of other issues (such as matrix interferences). Expand the discussion [at 11-20] to explain that only certain "J" values—those representing detections with concentrations likely below a quantitation limit—are to be treated "as the highest group of tied non-detects [sic]." Consider, too, a more precise use of language: it is potentially confusing to call these manifestly detected values "non-detects"!

There is a simpler method to adjust the Wilcoxon statistic for ties. I recommend presenting it. The formula 11.11 [at 11-21] was developed at a time when variances were painful to compute. It has been widely quoted ever since, despite enormous changes in computational capabilities. For hand computation, it is a shortcut, but for implementation in a computer program or spreadsheet, it's painful to carry out. It is much better—and far more intuitive—to provide a formula that works *whether or not any ties are present*; namely,

$$SD(W) = \sqrt{(Var(Ranks) * m * n / N)}.$$

Here, Var(Ranks) is the usual *unbiased* variance estimator of the ranks of the data. When the mid-rank convention is used for tied values, this expression is algebraically equivalent to the more complex expression in equation 11.11. With no tied values it reduces to formula 11.9 [at 11-18].

It would help to complete this section [at 11-21] by carrying out an example involving ties. (Perhaps it would be simpler not to provide a separate section on handling ties, so that one example will do. Instead, change the data in Example 11-3 [at 11-19] to include some nondetects and tied values. Describe the midrank convention in Procedure 11.4.3. Replace formula 11.9 [at 11-18] by the universal formula above. Move the discussion of "U", "E", and "J" values from section 11.4.4 into 11.4.1.)

Chapter 12

This chapter seems to contain more pitfalls for the reader than preceding chapters. The problems stem partly from its role of introducing prediction limits while limiting their application to comparisons to a single compliance well, which is an unrealistic situation. In part, though, the problems may be due to the exposition, which on occasion is imprecise and inconsistent.

I suggest announcing the limited role of this chapter right in its summary, Section 12.1 [at 12-1, top]. State that this chapter's procedures and examples are intended to illustrate basic computational techniques, but they are not intended for routine application at most RCRA sites.

Use consistent terminology

Please use consistent terminology. The haphazard variation in prepositions is particularly confusing. Are prediction limits "constructed on" background data [at 12-2, bottom], made "around a future" value [at 12-15, bottom], made "for a future" value [at 12-16, top], or made "on a future" value [at 12-19, top]? (I would like to suggest that a PL is *based on* and *calculated from* background data and that it is *designed for* and *compared to* a future value or statistic.) Is there a difference between "confidence," [at 12-7, *e.g.*], "probability," [at 12-2], "confidence level" [at 12-2, bottom], "level of statistical confidence" [at 12-6], "confidence probability," [at 12-11 and 12-12], and "statistical coverage" [at 12-12, top]? (Likely not.) Being consistent with terminology will help stave off confusion. This consistency should apply not just within Chapter 12, but throughout the UG.

Address problems with multiple comparisons

The UG must clarify the meaning of a prediction limit. The attempt is made, but it is incomplete: "Prediction intervals are constructed to contain, with a specified probability, the next one or more sample value(s) or sample statistic(s)…" [at 12-2]. The subtle idea

behind a prediction interval—one that is abused many times later in this chapter, by the way—concerns the nature of this "probability." Although the statement is correct, doubtless many users will understand the probability to be *conditional on* the background data used to construct the interval. But that is not the case, as later chapters indirectly indicate. The probability depends jointly on background and "future" values. It might be best to be technical as well as specific here. In all the intended applications, the background data consist of an n-tuple of random variables $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ and the future data are modeled by an m-tuple of random variables $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_m)$ where all components of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are identically and independently distributed according to some unknown probability law. That law determines the joint probability of $(\boldsymbol{X}, \boldsymbol{Y})$, which I will write F. The upper prediction limit is a statistic PL determined solely by $\boldsymbol{X}$, $PL = p(\boldsymbol{X})$, say, and the statistic that it is predicting is determined solely by the "future values" $\boldsymbol{Y}$, say $q = q(\boldsymbol{Y})$. The "specified probability" we are talking about is the joint probability $\text{Prob}_F(p(\boldsymbol{X}) \geq q(\boldsymbol{Y}))$. Somehow, the UG has to present the equivalent of this, either in similar terminology or through words alone, if that is deemed necessary. Then, it needs to point out, explicitly and clearly, that if we consider another set of "future" variables $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_m)$, then *there is no simple general relationship between* $\text{Prob}(p(\boldsymbol{X}) \geq q(\boldsymbol{Y}))$, $\text{Prob}(p(\boldsymbol{X}) \geq q(\boldsymbol{Z}))$, and $\text{Prob}(p(\boldsymbol{X}) \geq q(\boldsymbol{Y})$ and $p(\boldsymbol{X}) \geq q(\boldsymbol{Z}))$ (apart from the fact that the first two are obviously equal). Thus, unless one has specifically calculated the latter probability, *it is generally incorrect to re-apply a prediction limit based on fixed background values to more future data than it was designed for*. Both the false positive rate and the power can be altered in difficult-to-predict ways. (This is one reason why I believe it is incorrect to develop these prediction limits based on the number of comparisons in one year. The relevant false positive rate should be computed for the *entire* period in which a single prediction limit will remain in use and then, if necessary, adjusted to a one-year equivalent for comparison with other statistical methods.)

Please change the examples to reflect this statistical truth. Most of them present an incorrect application of a prediction limit. Example 12-1 [at 12-9] computes a 95% prediction limit for *one group* of four future values but then applies it to *two groups*. Example 12-3 [at 12-21] computes a 99% prediction limit for *one mean* of four future values but then applies it to *three groups*. Example 12-4 [at 12-26] computes a 99.1% prediction limit for the median of *one group* of three future values but then applies it to *two groups*. *All* of these procedures will *exceed* their nominal false positive rates, many by more than most people would expect.

For instance, one could construe Example 12-1 as an instance of a 95% prediction limit applied in "year 4" followed by a verification sample obtained in "year 5." As such, the expectation created by the current language in the UG would be that the false positive rate is 0.05 * 0.05 = 0.25%. In fact, though, simulation suggests the false positive rate is 0.78% ± 0.028%: over three times greater than expected. Computing a prediction limit to cover all eight future values would decrease the power of this test somewhat (from 93% at 3 SDs to 84%), but it would still exceed the EPA reference power curve.

These considerations bring forward a key issue: why does the UG make such specific and strong recommendations concerning various levels of confidence and other aspects of the tests to use (such as limiting the number of future values to the quantity obtained between two successive tests [at 12-7])? It would be far better for the UG to describe and illustrate the processes of (a) checking the appropriateness of the test and (b) demonstrating that appropriate detection power has been achieved. In Example 12-1, for instance, the UG could illustrate the relevant power calculations, instead of misleadingly applying the same PL twice to the two successive years of data.

Avoid vagueness

One particularly confusing aspect of this chapter, as well as of many of those that precede it, is its tendency to be vague about the intended applications. Take, for example, the "Requirements and Assumptions" section, 12.3.2 [at 12-7]. Its first paragraph is sufficiently general—especially when we take "background" in the more general sense of being a collection of reference data (obtained either from upgradient wells or at earlier times from a compliance well) —that one could fairly suppose that it applies both to interwell and intrawell testing. The next paragraph, though, shakes this assumption. If "a new prediction interval should be constructed" for "each successive evaluation period," then are we limiting the discussion to interwell comparisons only? Or perhaps is the UG allowing for the routine updating of an intrawell background dataset (a suspect procedure, but one that many readers might attempt, especially in light of the subsequent remark that "background data should be amassed or accumulated over time" [at 12-8, top]). And then exactly how are "Welch's t-test or the Wilcoxon rank-sum procedure" [at 12-8, top] supposed to be applied to obtain "evidence of characteristic changes within the background groundwater quality"? At what level of significance?

Overall, this material provides useful advice to the experienced statistician, who can make educated guesses about its intended application, but it is so vague about the specific applications that I believe we cannot expect the intended audience to interpret it accurately or reliably. In order to reach this audience, the UG must be consistently and explicitly clear about

(a) What constitutes "background" data in any intended application.

(b) What constitutes the "compliance" data.

(c) How frequently the background data should be updated.

(d) How frequently diagnostic tests (of Normality, heteroscedasticity, etc.) should routinely be performed.

(e) What levels of significance are recommended for the formal tests and the diagnostic tests.

(f) What constitutes "verification resamples."

(g) How exactly any verification resamples should be incorporated in the testing.

(h) The extent to which verification sampling should be formally included within the test description and computations of significance and power.

Speaking of this audience, it strikes me that equations such as 12.5 [at 12-12], 12.7, and 12.9 [at 12-16], will not be at all meaningful or useful. Put these in a footnote or appendix.

Use notation consistently

The UG uses the letter *j* as a "rank" for which $j = n$ is the maximum value [at 12-12]. Later, in a similar context [at 12-24], it uses the same letter *j* as an "order statistic" where now $j = 1$ is the maximum value! This is likely to cause confusion and errors. Please use consistent terminology and notation in these sections (and in other chapters, where appropriate). It is more important to be internally consistent with notation than it is to agree with the referenced papers.

Disallow dynamic (*post-hoc*) test selection

Up to this point (at the end of Chapter 12), the UG has not clearly distinguished the process of test selection from test execution. In many examples the UG specifically recommends changing a test, sometimes even changing what the test does (e.g., from comparing means to comparing medians), as part of its very application. This is very troubling.

Example 12-4 [at 12-26] illustrates the problem nicely. This example presents a background (upgradient) set of 24 data and two sets of compliance well data, three values per well. Departing from the example, let's suppose that the reader's original intention is to use a nonparametric prediction limit for the largest of three future values. The PL is computed as the maximum background value, 9.2. By equation 12.6 [at 12-12], its confidence is 24/(24+3) = about 89%. In applying the test, though, the reader will see that the largest value at one compliance well of 10.4 exceeds this PL. Not wanting to report this, the reader then turns around and uses the background maximum as a 99.1% PL of the *median* of three future samples, as in the example. This time, no median compliance value exceeds 9.2 and there is nothing to report.

This scenario is not fanciful at all: owners, operators, and their consultants want to comply with the regulations and their permits. Since the UG seems to allow any test to be performed, almost willy-nilly (especially non-parametric ones), one can expect a lot of effort to go into *finding some* post-hoc *test that makes the site look clean.* The defense against that form of abuse, of course, is to require that the permit specify exactly what test will be followed. A well-written permit can include recommendations from the UG, stating

when and under what conditions data might be transformed to Normality, for instance, or when one test might be chosen over another. The crucial point is that all this must be specified in advance. I do not believe the UG has made this point. This is worth an extended discussion, perhaps in Section I. In particular, provide explicit guidance concerning what information should go into the permit itself concerning the execution of the statistical tests, their diagnostics, and choices for alternative tests.

Chapter 13

Portions of this chapter (13.1 through 13.4), along with parts of Chapter 6, contain the best written material in the UG. It is clear, thoughtful, helpful.

Organization

Sections 13.1 through 13.4 (summary, background, historical notes, and "Computing … FWFPR") as well as parts of 13.8 ("Sites Using More Than One Statistical Method" and "Hypothetical Case Studies") contain material that is more general than the putative subject of this chapter, PLs for multiple comparisons. Principles of logical organization would suggest moving this material into Section I.

Suggestions for readability

Some minor clarifications would improve the already good discussion in sections 13.1 through 13.4:

- Provide a clear, precise definition of "evaluation" [at 13-5, *e.g.*].
- Clarify what it means for a "chemical" to "equal … the QL or RL" [at 13-9, bottom]. What happens, for instance, if historical QLs are 10 mg/L but then the laboratory begins using a QL of 1 mg/L? Is a quantified value of 2 mg/L above or below "the QL"?
- Resolve the discrepancy between the guidance to have high effective power "three or more standard deviations above the background mean" [at 13-15, bottom] and "two or more standard deviations above background" [at 13-15, footnote 4].
- Indicate [at 13-19] that the Bonferroni adjustment $\alpha^* = \alpha/n$ is an approximation to the more accurate expression $\alpha^* = 1 - (1-\alpha)^{1/n}$. The correct value is approximately $1 + \alpha/2$ times greater than the approximation (because the limiting value of the ratio is $-\ln(1-\alpha)/\alpha$), showing that it is gets poorer as α gets larger. Acknowledging that this is an approximation would make it easier for readers to use the more accurate calculation.

- Explain the sense in which retesting is "an alternate strategy" [at 13-22]. Alternat[ive] to what? What primary strategy is implied? Why is retesting being deprecated?

- Consider using a latin letter rather than ω (Greek omega) [at 13-29], because many readers will not recognize or be comfortable with Greek letters. In some places it should be ok to retain conventional Greek letters, mainly μ and σ, but there it would be helpful to provide their names in parentheses, as in "…with a mean μ (Greek "mu") shifted upward…" [at 13-73, top].

Points that might cause confusion

Sections 13.5 through 13.8 are more problematic than the preceding ones. Some specific issues to look into are:

- Example 13-3 [at 13-37 and 13-38] potentially confuses as much as it clarifies the point. The data show apparent trends over time; they are gathered at very uneven intervals over time; some intervals are within just a few weeks of each other, raising the possibility of serial correlation influences; the logarithms are mysteriously shown with four or five more significant digits than the original concentrations (why?). In general, use "clean," simple datasets for examples so that the method is clearly illustrated without needlessly creating possible objections.

- Acknowledge potential holding time problems with resampling schemes. For instance, in footnote 7 [at 13-38], it simply is not possible to collect potential resamples and hold them to the end of an evaluation period (often up to six months or a year) prior to chemical analysis.

- Parenthesize equations properly. For instance, the expression "$0.10/c \bullet n_E$" appears frequently [as at 13-49]. According to conventional rules of algebraic precedence, that ought to be interpreted as equivalent to $(0.10/c) \bullet n_E$, but that is not what is intended.

- Clearly distinguish steps to carry out during permit development (that is, test selection) from those that are carried out routinely during test evaluation. For instance, the first step of Procedure 13.6.2.2 [at 13-52] is to "check first for normality" of the historical observations. This would be done during permit development, but does not have to be performed each time the test is conducted.

- Clarify the distinction being made between "chemical" [as at 13-8, top, 13-9, bottom, and 13-38, top], "parameter" [at 13-3, 13-6, etc.] and "indicator parameter" [first appearing at 13-2]. In particular, "indicator parameter" has a specific conventional meaning in RCRA groundwater monitoring that

appears to differ from the use here: indicator parameters typically do not measure specific chemicals at all and often are limited to pH, specific conductance, TOC, and TDS.

- Rewrite the paragraph spanning pages 13-57 and 13-58. This is utterly confusing. Its opening line asserts it intends to discuss the power of "non-parametric retesting schemes," but it proceeds instead to discuss "parametric intervals." Its resulting conclusions are impossible to interpret with any confidence.

- Explain the sense in which "κ factors" are *not* "determined solely from the background measurements themselves" [at 13-58].

- Choose appropriate verbs and prepositions. For instance, the word "delineate" [at 13-71, bottom] cannot be the intended verb in the sentence "So each distinct data configuration and retesting plan … would delineate a different statistical test method." Perhaps "determine" is meant here? *Strange choices of verbs and inconsistent use of prepositions* are problems that plague many parts of the UG, not just this chapter. For example, what distinction is being implied by the change of prepositions in "… plans involving prediction limits *on* means tend to be more powerful than similar plans using prediction limits *for* observations" [at 13-66, emphases added]?

- Separate example statements from their solutions. For instance, the statement of Example 13-7 [at 13-63] does not assert that "a single evaluation is done annually;" somehow, the reader is supposed to know that or guess it. The UG tells us this fact only in Step 1 of the *solution* [at 13-63]. Example 13-9 [at 13-74] does not state that the "inorganic constituents" will be naturally-occurring. The solution *assumes* this. (Although *all* inorganic compounds will occur naturally at some concentration, many of them, such as mercury, silver, thallium, selenium, cadmium, and hexavalent chromium, are commonly not detectable using standard EPA techniques.) For the sake of developing clear, unambiguous examples, it would be better to make such assumptions in the statement of the problem rather than the description of its solution.

- Avoid the use of derogatory colloquialisms. Phrases like "contaminated culprit" [at 13-70] overtly suggest a bias on the part of the UG authors and are likely to offend owners and operators.

Provide statistical details

Chapter 13 is strikingly different from all other chapters of this guidance and from all preceding guidance in that it relies entirely on tables and software *without providing any method to calculate the tabulated values*, to verify whether they are correct, or to extend them to areas not covered by the tables. Why are computational methods not provided? I

recommend completing this chapter by providing explicit algorithms for reproducing all entries in all the tables, even if that might require technical details best placed in an Appendix. Any reader who can correctly carry out the Monte-Carlo calculations of power [at 13-73, *e.g.*] surely can implement any of the computations needed to create these tables.

3.IV. For Section IV of the Unified Guidance, please address the following questions :

a. Does Section IV meet the stated objectives described in the Introduction to the Charge, above? Please explain.

It does a good job, subject to the proviso that much of the research on control charts that it announces is likely to produce methods that supersede this guidance.

b. Does Section IV cover an appropriate range of topics? Are there any key topics that are missing or that should be emphasized or described in further detail? Please explain.

Chapter 15

Describe how to interpolate values in Table 15-1. Gilbert recommends cubic interpolation, which is somewhat complicated and requires guidance and an example. A simpler method is bilinear interpolation of the values of $z = \sqrt{[n \ln(H/s_y)]}$ (so that $H = s_y \exp(z^2/n)$. For example, either interpolation method obtains $H = 4.090$ instead of the 4.069 given in Example 15-3 [at 15-15]. The resulting UCL is 18.9 ppb rather than 18.7 ppb.

Show how to use the Normal approximation to the Binomial (with continuity correction) to obtain confidence limits of the median when $n > 20$ [at 15-19].

Recommend plotting nondetects and other unquantified values using symbols that differentiate them from quantified values [at 15-36]. In this fashion, one can assess the extent to which apparent trends may be due to changes in the censoring limits. In such plots, it is best to use the censoring limit to plot the value.

c. Is the material in Section IV organized and presented in a clear and concise manner? Please explain.

Chapter 15

This chapter, although it contains good discussions and excellent advice, appears to have been edited with less care than preceding chapters: it contains many confusing sections and plenty of technical errors.

The potential confusion begins with the chapter title: although the topic is confidence *limits*, the title is about *intervals*. Because the text does not clearly explain the distinction between two-sided and one-sided intervals and their relationship to limits, this is a potential problem.

Figure 15-1 [at 15-4] will be difficult to interpret until axes appear, with labels. At a minimum, inform the reader that the upwards direction corresponds to increasing concentration and the direction to the right corresponds to increasing time.

The statement about the relationship between confidence levels and significance levels [at 15-5] is correct only for one-sided tests. Rectify this by discussing two-sided intervals and tests in detail.

Procedure 15.2.3 [at 15-6] is difficult to follow because a context has not been provided. The language alternately refers to "*the* compliance well" and "*each* well" (emphasis added). So: is the procedure designed for a set of compliance wells or just one? Are there other circumstances in which it might be applied.? Clarify this by stating exactly what the monitoring context is.

To make Procedure 15.2.3 clear and self-contained, provide explicit formulae for the mean and standard deviation.

In Step 2 of Procedure 15.2.3 [at 15-7, top], clarify the meaning of the t-statistic by stating that it is the upper 1-α percentile of Student's t distribution, rather than vaguely stating that it is "obtained from a Student's t-table."

Modify the incorrect conclusion in Step 3 of Procedure 15.2.3 [at 15-7]. When a confidence limit exceeds a standard, one should *not* necessarily conclude "that further corrective action is needed." An alternative conclusion is that additional data should be collected. Rectify similar lapses in Procedure 15.3.3 [at 15-11, bottom] and Example 15-5 [at 15-27]. (Procedure 15.4.3, in contrast, makes the more supportable conclusion that "there is insufficient evidence that the clean-up target has been achieved" [at 15-15].)

Clarify Example 15-1 by stating that the significance level will apply separately at each well, rather than being a facility-wide level [at 15-7, bottom].

Adopt a consistent terminology. Is the subject of this chapter a confidence interval, limit, bound, or target? (All four terms appear, apparently synonymously, within a single paragraph [at 15-18].) The inconsistency is particularly glaring in Example 15-1 where the abbreviation for "lower confidence interval bound" [sic] is given as "LCL" [at 15-8]!

Please explain what is meant by "a comparison of the GWPS … will provide a more reasonable test of long-term exposures" [at 15-10]. What, exactly, is the intended relationship between *concentrations in groundwater samples* and (undefined) *exposures*? In human health and ecological risk assessment, there is a relationship, but it certainly is not as direct as implied by this statement, and it also varies tremendously from one site to another. Consider limiting the discussion to groundwater concentrations, leaving considerations of exposure to risk assessment guidance.

Provide formulas for the log mean and log SD in Procedure 15.3.3 [at 15-11]. This is especially important because it is not immediately clear that one ought to use the usual estimator of SD that incorporates a bias-correction term.

Remove the "sample number" column from Example 15-2 [at 15-12]. It serves no purpose, and so can only create possible confusion or misinterpretation.

For methods of dealing with the nondetect in Example 15-2 [at 15-12], refer explicitly to Section 10.3. It would be worthwhile to discuss how to perform a sensitivity analysis of the imputation of the nondetect value, rather than proposing one-half the detection limit as a "reasonable compromise." For instance, if one varies the imputed value between zero and the detection limit, the UCL is minimized at an imputed concentration of 0.23009 (where the UCL attains a value of 2.8457); it cannot be any lower using a simple substitution method. Using Cohen's Method, I obtain 2.727 for the UCL.

In the examples, indicate how all numbers were computed. For instance, in Step 5 of Example 15-3 [at 15-16], indicate that the CV of 1.965 is $\sqrt{(\exp(s_y^2) - 1)} = \sqrt{(\exp(1.2575^2) - 1)}$.

Simplify the procedures for obtaining confidence intervals around medians. In general, the reader will only be computing one-sided intervals. Step 2 of Procedure 15.5.3 [at 15-19] requires a complicated, confusing process to generate a symmetric two-sided interval. This is completely unnecessary in the one-sided application. Omit this complication, or relegate it to a separate procedure (or footnote).

Resolve the inconsistency between guidance to "approximately" round actual confidence levels [at 15-21, bottom] and to obtain actual levels that do not exceed the nominal levels [at 15-32, Step 3]. If some rounding procedure is still recommended, then provide explicit rules for how to round and when.

Clarify the discussion of non-parametric confidence limits around upper percentiles: How can it be the case that "the next best choice is a confidence interval around an upper percentile close to the maximum" [at 15-28, bottom] when "there is no maximum value associated with continuous distributions" [at 15-23]?

Recommend plotting residuals against *estimated* concentrations instead of plotting residuals against concentrations [at 15-34]. Figure 15-5 [at 15-40] illustrates the former rather than the latter.

4. Are the methods, approaches, and strategies described in the Unified Guidance technically valid and accurately interpreted, described, and applied in a groundwater monitoring context? Please comment on specific methods, approaches, and strategies, as appropriate.

To a great degree, material in the UG is technically valid, accurate, and applied correctly and clearly. Examples appear realistic and well chosen. Typographical errors are practically nonexistent. In most chapters, numerical errors are difficult to find. It is clear that a great deal of attention has been paid to accuracy in the examples and test descriptions.

Chapter 6

Some software computes a "skewness" different from that presented here[5] [at 6-19]. The UG should warn readers about the difference and provide guidance (such as to recommend that readers learn exactly what their software is calculating and adjust its output as necessary).

At [6-2], "absorption" should be "adsorption."

The Central Limit Theorem is abused here [at 6-4]. It would be better to say "sums of independent random quantities of similar variances tend to follow a Normal distribution."

Is the UG [at 6-10, top] recommending *routine* diagnostic testing (with every sample event), diagnostic testing at planned intervals (such as at two-year evaluations), or only during test selection (during the permit development process)?

Chapter 8

I disagree with the assertion [at 8-13] that "normality is not a bad default assumption for these napthalene [sic] data." The probability plots clearly show strong departures from normality. Perhaps the best response would be to modify the example data so that they do not depart so obviously from normality. As a matter of pedagogy and communication, it could be distracting to retain this potentially controversial (but not terribly relevant) aspect of the example in question.

---

[5] Systat and Excel, for instance, in order to make the coefficient an unbiased estimator of population skewness, replace the factor of $\sqrt{[n/(n\text{-}1)^3]}$ by $n/[(n\text{-}1)(n\text{-}2)]$. The ratio, $\sqrt{[n(n\text{-}1)]} / (n\text{-}2)$, is small but appreciable: in Example 6-1 [at 6-20], Excel computes a skewness of 2.00 rather than 1.84. To appreciate the effect, we can approximate this ratio by replacing the geometric mean of n and n-1 by the arithmetic mean, giving $(n\text{-}½)/(n\text{-}2) = 1 + 3/(2n\text{-}4)$. Evidently the ratio is largest for small n, with values 1.73, 1.25, 1.11, and 1.05 at n = 4, 8, 16, and 32, respectively.

The illustration of Rosner's test [at 8-16] could be improved in several ways. The table for Step 3 does not conform to the procedure given [at 8-12]. According to the procedure, only two columns of calculations are needed: the third (rightmost one) is unnecessary. Remove it. To make the illustration complete, add a line to this table showing the values of R_0 and R_1.

The illustration would work better if it also showed a slightly more complicated example of Rosner's test. For instance, changing the value for Well 3 in Quarter 3 in Example 8-3 [at 8-13] from 23.23 to a value just less than 12.34 would cause it still to look like a possible outlier on the Normal probability plot, but not to be an outlier at the 5% significance level. The test would still identify the extreme value of 35.45 as an outlier, thereby illustrating the sequential search for outliers within the original block of two. Doing this would obviate the need for the note [at 8-16].

Table 8-2 (in the Appendices) needs a caption explaining its use. By itself, it is unclear how one uses the multiple entries that appear. It is only by closely following Step 4 of Example 8-3 [at 8-16] that one can determine how this table is to be read. By giving explicit instructions, fewer errors will result.

At several places in Chapter 8, the UG recommends "deleting" an observation [at 8-10, top, for instance]. It would help to be specific about what constitutes "deletion." I would like to suggest that this procedure be strictly a *temporary* one, specific to a statistical test conducted under the RCRA regulations. In particular, absent convincing evidence that the observation is a genuine error, it should remain in the monitoring database (perhaps flagged as an outlier) and, in many cases, it should continue to be shown in graphical representations and tables of the data, visibly identified as an outlier.

Chapter 9

Typographical issues

Example 9-1 [at 9-12] contains a typographical error: the median value for Well 1 is 50.06, not 55.06.

I cannot reproduce the computations of Example 9-3 [at 9-23, bottom]. For the 99% PLs I obtain 171.4, 198.2, 290.6, 247.6, 458.0, and 558.0. (Whether I use the rounded logs displayed [at 9-19] or the more accurate logs computed from the data [at 9-13] sometimes affects the least significant digit by one, so this cannot explain the differences.) Part of the discrepancy stems from a computation shown in Example 9-2, Step 2 [at 9-20]: I obtain 4.331 for SS_{wells} rather than 4.294. Consequently I obtain different sums of squares, mean squares, and F-statistic in the ANOVA [at 9-20]. This is despite being able to reproduce the means and SDs shown earlier [at 9-19, bottom] to acceptable accuracy, indicating that I have not made a data transcription error. Please review these calculations.

Problems with the ANOVA-based prediction limit

Example 9-3 [at 9-23, Step 3] introduces a prediction limit test that is not described or referenced. The test sometimes is invalid: it can result in false positive rates substantially greater (twice as great, for instance) than the nominal rate (alpha).

To demonstrate this assertion, let's ignore the unnecessary complication of the log transformation. The test effectively states that a prediction limit for a single future value Y_0 at a well given by

$$PL_{1-\alpha} = \overline{Y} + t_{1-\alpha,df}\sqrt{MS_{error}\left(1+\frac{1}{df+1}\right)}$$

where $\overline{Y}$ is the mean of n observations *at that well only*. "*Df*" equals 18 in the example and evidently is $w(n-1)$ in general, the number of wells times one less than the number of observations per well [*cf* 9-22, middle]. Let the true mean at that well be μ and let the common variance (for all wells) be σ^2. Three independent random variables appear in the prediction limit expression:

$$Y_0 \sim N(\mu, \sigma^2),$$
$$\overline{Y} \sim N(\mu, \sigma^2 / n), \text{ and}$$
$$MS_{error} \sim \chi^2(df)\sigma^2 / df .$$

The variance of $\overline{Y} - Y_0$ evidently is $\sigma^2(1 + 1/n)$, implying that $\dfrac{\overline{Y} - Y_0}{\sqrt{MS_{error}\left(1+\frac{1}{n}\right)}}$ has a Student t distribution with df degrees of freedom (because MS_{error} is estimated with df degrees of freedom). Therefore the correct prediction limit would appear to be

$$PL'_{1-\alpha} = \overline{Y} + t_{1-\alpha,df}\sqrt{MS_{error}\left(1+\frac{1}{n}\right)},$$

which always is a little bit larger than the formula given by the UG.

However, even this corrected formula has problems, as simulations show. For instance, I conducted one simulation (with 10,000 iterations) that mimics the data of Example 9-3 [shown at 9-19]. By rescaling, we lose no generality in assuming the common distribution of the iid errors is Standard Normal. In this particular simulation $\alpha = 0.99$ (as in the example). I set the expectations of the six wells to -1, -0.6, -0.2, 0.2, 0.6, and 1 (that is, uniformly spaced from -1 to 1). In each iteration of the simulation, four iid observations were created for each well and an additional independent observation (representing the future value y_0) with the same expectation was also created for each

well, resulting in 6*4 + 6 = 30 simulated observations in all. In 5,799 of the iterations, the F-statistic (based on 24 observations) was significant at the 5% level (as indicated by the UG [at 9-19, Step 9]). At Well 1, in 5,703 of these cases the PL computed using the *correct* formula above actually covered the simulated future observation for Well 1. The corresponding counts for the other five wells were 5,721, 5,712, 5,728, 5,724, and 5,732. As proportions of the 5,799 "significant" iterations, the six values are 98.34%, 98.65%, 98.50%, 98.78%, 98.71%, and 98.84%. It appears there is an approximate linear relationship between the coverage frequency and underlying mean in this case (P = 3.4%). Nevertheless, in this simulation *all* proportions were less than the expected value of 99% (the nominal α). Their mean, 98.64%, is far enough below 99% to suggest this phenomenon is repeatable, not a simulation artifact[6]. The observed false positive rate therefore is (100 – 98.64) / (100 – 99) = 1.36 times greater than the nominal false positive rate. This is not bad, but it is possible that in other situations (with different values of w, n, α, and well expectations) the difference could be greater. Using the incorrect formula given in the UG, the false positive rates can easily double the nominal rates.

Intuitively, what's happening here is that screening with the F-test winnows out those situations where variation is toward the high side at the low-mean wells and toward the low side at the high-mean wells, because those situations are less likely to be found "significant." Consequently, there will be a tendency for the estimated means to be more spread out than they really are. (In the simulation, among the "significant" iterations the estimated well means were typically 16% further from zero than the true underlying means, while among all iterations the estimated well means agreed with the true underlying means.) As a result, this creates a tendency for the PL of a low-mean well to be underestimated and the PL of a high-mean well to be overestimated. Because the estimated means are too spread out, the between-wells variance is overestimated, resulting in underestimation of the error variance. (In the simulation, the mean value of $\sqrt{MS_{error}}$ among the "significant" iterations was 0.92, eight percent lower than the underlying value of 1.) This causes all PLs to be underestimated. The combined effect of biased mean estimation and underestimation of the SD (at least in this one case) is to underestimate all PLs, with the lower-mean wells exhibiting a greater tendency toward underestimation.

I therefore cannot concur with any recommendation in the UG to use ANOVA estimates of standard deviations for constructing *intrawell* prediction limits. This approach looks like it may work in some cases, but until those cases are accurately characterized, readers risk experiencing substantially higher false positive rates than anticipated.

[6] Two repetitions of this simulation produced comparable results. The mean false positive rates were 1.50% and 1.40%. In another simulation with well means ranging uniformly from -0.5 to +0.5, the false positive rate for the PLs ranged from 3.34% down to 1.34%, all substantially greater than the nominal 1%. (In this simulation of 40,000 iterations, 6504 (16.26%) were significant.)

It may seem that distinguishing a false positive rate of 1.5% from 1% is debating a rather fine point. Considering, however, that a proportional increase in the individual false positive rates will create (approximately) the same proportional increase in the facility-wide false positive rate, we see that a target FWFPR of 10% might really be 15% or greater, depending on the circumstances. Small changes in false positive rates can have an appreciable effect.

One might counter that the same kind of consideration ought to be given to the power of these tests. I agree. Maintaining power is critical. However, there is an important phenomenon assisting us here: the repetition of these tests over a sequence of monitoring events improves the power (while making the false positive rate worse and worse). Thus, at least if a facility does not have an environmental need to detect and respond to a release within the span of one monitoring period, then it is highly likely that a false negative will be followed by a true positive result in short order. This indicates that accurate and precise assessment of false positive rates is critical.

Another possible objection is that looking only at the "significant" results in a simulation is what creates these biases in the first place. That is true. The non-significant results mimic potentially real situations where there are true underlying differences in well means, but the ANOVA does not detect these. The facility might in such cases proceed to pool all data, thereby overestimating the common variance, and use a common mean (rather than individual well means) as the basis for a single, sitewide PL. This PL would generally be larger than desired for most individual wells. The false positive rate would decrease below its nominal value and the power would decrease. In reality, though, the moment at which a test is selected occurs *after* the ANOVA is run, not before, implying that the correct operating characteristic to compute is the one *conditional on* observing a significant F-statistic.: the ANOVA-based PL simply is never computed otherwise. Test selection and test performance interact in a very interesting way.

Chapter 11

The results of Example 11-2 [at 11-15] appear to be incorrect. In general form the trend of the powers is ok, but the values are accurate only to the first significant digit. The correct powers for $k = 0.5[0.5]5$ are, to four decimal places, 0.1646, 0.3580, 0.5975, 0.8041, 0.9285, 0.9808, 0.9963, 0.9995, 0.9999, and 1.0000. Please provide the correct results[7]. If the incorrect ones were created using software recommended in the UG, then please warn readers and change the recommendation.

[7] I computed these using Algorithm AS 243 Appl. Statist. (1989), vol.38, no. 1: Cumulative probability at T of the non-central t-distribution with DF degrees of freedom and non-centrality parameter δ. This is publicly available as an Excel macro ("NCTDist") at http://www.quantdec.com/envstats/software/intervals.xls. (The same macro exactly reproduced all of Table 15-2, "Factors (κ) for parametric upper confidence bounds on percentiles (P)" [at C-186 through C-189].) I confirmed several of these results using the online statistical calculator from the UCLA department of statistics at http://calculators.stat.ucla.edu/cdf/ncstudent/ncstudentcalc.php, obtaining agreement to at least seven significant figures.

Chapter 12

Remove the Poisson PL from the UG

I can find no legitimate basis to recommend this test. In general, the justification for any test will rest on one or both of two assertions: either (a) theoretical considerations suggest it will be suitable or (b) experience shows it just plain works.

The Poisson PL has no theoretical justification that has stood up to scrutiny. Gibbons originally argued that concentrations reflected counts. This might be true—it was a clever suggestion—but the original counts are long lost by the time concentrations have been reported. We simply don't know whether one ppb corresponds to a count of 1, 10, 100, or something else on an instrument. (I suspect that the correspondence likely changes from one time to the next, depending on the instrument, its calibration, any dilutions of the sample, and so on.) The prescription for rescaling concentrations [at 12-29 and 12-30] is an unjustified *ad-hoc* attempt to recover these unknown counts. Unfortunately, *the test is very sensitive to the choice of scale factor.*

Furthermore, the proposed treatment of nondetects (at [12-30, top]) clearly makes the data non-Poisson: zeros cannot possibly appear. One should at least use a left-truncated Poisson model rather than the Poisson distribution itself. Again, *the test is very sensitive to this choice.*

If the Poisson PL is to be recommended, it can solely be on the basis of its efficacy. That it cannot work reliably, though, is amply illustrated by some simple simulations. Emulating Example 12-5 [at 12-31], I created a simulation wherein six iid background Poisson variates were generated and three more independent "compliance" data were generated from a Poisson distribution having mean equal to the background, the background plus two standard deviations, and the background plus four SDs. These Poisson counts were then converted into concentrations using a fixed scale factor, left-censored at a fixed detection limit, and processed strictly according to procedure 12.7.3 [at 12-30]. Specifically, a scale factor was determined from the background observations; all values (background and compliance) were rescaled accordingly; and the nondetects were replaced by one-half their rescaled detection limits. The rescaled compliance well values were then compared to the Poisson PL based on background.

During the simulations I preset the detection limit in order to obtain about 70% to 90+% nondetects in background on the average, because this is the situation in which "the Poisson model is sometimes justified" [at 12-27]. The simulations varied primarily in the Poisson mean (a count, not a concentration). With large means (2-20), the test simply has no power to detect even a four SD increase. With moderate means (1), it achieves results comparable to those reported in Chapter 10 [at 10-9 and 10-10]. With smaller means (theoretically possible: remember, since there's no valid physical justification, we're just trying to find something that might work), the power increases but the false positive rate balloons. In short, *I could find no parameter for the underlying*

Poisson distribution that gives this test any simultaneously acceptable combination of significance and power levels.

This test, therefore, has everything going against it: it has no valid justification, it is sensitive to how nondetects are treated and how the concentrations are rescaled, and it cannot balance the false positive and false negative error rates. It cannot meet the RCRA standards.

Check the calculations

The mean of the background well logs in Example 12-3 [at 12-22, top] is 2.553, not 2.533. The resulting PL is 3.85 rather than 3.83.

Chapter 14

Sen's slope estimator is incorrectly described and applied [at 14-31 and 14-32]. The correct pairwise slope between observation (t_i, x_i) and (t_j, x_j) (where the times are different) is $(x_j - x_i) / (t_j - t_i)$. This is clear enough in Gilbert's formulation (pp 217-218), especially when one examines his computer code. The formula originally appeared in (Sen 1968) at equation 3.1.

Using the correct formula, Sen's slope estimate for the data in Example 14-4 [at 14-31] is 1.333 ppm/yr rather than "almost 2 ppm [per] … year" [at 14-31]. Sen also provided a method to estimate confidence limits around the slope; the computations are no more difficult than those associated with the Mann-Kendall estimator and therefore could be described and recommended in the UG. (If we were to pretend that the slope estimator and the median of the x's were independent, then conceivably we could construct a non-parametric version of the confidence bands around the fitted line described in Chapter 15 [at 15-36 and 15-37]. I am not aware that anyone has done this. Future research?)

Chapter 15

Fix the errors in Table 15-1 (Land's H factors) [at C-170 through C-185]. About one percent of the values are incorrect[8]. I found these errors by estimating the second mixed partial derivative of H (with respect to n and s_y) and searching for patterns characteristic of isolated errors and then iterating this procedure after fixing the initial errors. I did not compute any values *ab initio*, but only modified the erroneous values until they fit within the context of the neighboring values in the table. Consequently, my check was exhaustive, except for the smallest values of n (3) and s_y (0.10), and my corrected values might be off in the last decimal place.

[8] The errors appear to be of two types. In some places, negative signs are omitted from swathes of values. This looks like an optical character recognition problem. In other places, the digits of isolated values are sporadically changed or transposed; sometimes a value from one table is put into the same location in a different table. These are errors a human transcriber would make. This all suggests that the errors in Table 15-1 may have been introduced in different ways at different times; some of them might even occur in Land's original publication.

Table	*n*	s_y	H (corrected)
$H_{0.01}$	4	0.20	-3.089
$H_{0.01}$	13	0.30	Insert "-"
$H_{0.01}$	15	1.50	Insert "-"
$H_{0.01}$	25	0.30	Insert "-"
$H_{0.025}$	7	0.80	-1.882
$H_{0.025}$	15	1.75	Insert "-"
$H_{0.025}$	17	1.00	-2.019
$H_{0.025}$	19	1.00	-2.036
$H_{0.025}$	21	0.70	-1.966
$H_{0.025}$	25	0.90	-2.033
$H_{0.05}$	14	9.00	-6.908
$H_{0.05}$	17	3.00 – 10.00	Insert "-"
$H_{0.05}$	28	2.00	-2.296
$H_{0.05}$	31	10.00	-8.756
$H_{0.05}$	36	0.70	-1.694
$H_{0.10}$	13	1.25	-1.417
$H_{0.10}$	23	8.00	-5.502
$H_{0.10}$	25	7.00	-4.904
$H_{0.10}$	31	10.00	-7.090
$H_{0.10}$	11	2.00	Insert "-"
$H_{0.10}$	17	3.00 – 10.00	Insert "-"
$H_{0.90}$	9	0.70	1.840
$H_{0.90}$	13	1.75	3.019
$H_{0.90}$	14	4.00	6.229
$H_{0.90}$	16	0.70	1.677
$H_{0.95}$	31	10.00	17.13
$H_{0.975}$	9	0.10	2.281
$H_{0.975}$	25	4.50	10.51
$H_{0.975}$	31	10.00	21.64
$H_{0.99}$	6	0.90	8.586
$H_{0.99}$	12	0.80	4.411
$H_{0.99}$	12	1.50	7.012
$H_{0.99}$	14	3.00	12.01
$H_{0.99}$	21	4.50	14.54
$H_{0.99}$	28	10.00	28.62
$H_{0.99}$	31	10.00	27.73

It would help to extend these tables, at least to $n = 101$, as in Gilbert's book (1987).

Fix the errors in Example 15-7 [at 15-40, top]. According to equations 15.24 and 15.25 [at 15-37], the value in the denominator of the second fraction is n-1, not n-2. Thus, the "8" should be a "9". The resulting UCLs are 12.87and 18.14 ppb.

5. In your opinion, what are the weakest and strongest aspects of the various sections, chapters and/or recommended methods? Please make suggestions on how the weakest parts can be strengthened.

Section I

In Section I, Chapter 3 (the overview) stands out as particularly clear and useful. It is thorough and clear. It looks like it got a lot of attention and review.

In Section I, Chapter 4 (statistical background) appears weaker by comparison. The problem lies not with the material or its exegesis, but rather in a mis-match with the intended audience. In its effort to reach the statistically less sophisticated reader, it could be perceived in some places as sloppy and patronizing.

The potential to appear sloppy comes about through the misuse of some words and the failure to define others carefully. Take, for instance, the assertion that "… certain steps are involved in conducting any statistical hypothesis test. … First, the null hypothesis must be established" [at 4-2, bottom]. A word that better reflects the meaning that must have been intended would be "specified" rather than "established."

Later in the same paragraph, the UG writes that "the observed data … is assumed to follow a known statistical distribution…" What, exactly, does it mean to be "known"? In almost all applications, in fact the underlying distribution is not known. We assume only that it is one (*unknown*) member of a family, such as the set of Normal distributions of arbitrary variance. The statistician, intimately familiar with this situation, will of course understand what is intended, but it seems too much to hope that even a clear-thinking hydrogeologist or EPA case manager will decipher this correctly.

This same paragraph suggests that an alternative hypothesis might be of the form H_A: [Benzene] ~ Normal(20 ppb, σ^2) where the null hypothesis is H_A: [Benzene] ~ Normal(0 ppb, τ^2). This is exceptionally unlikely and unrealistic. Usually the alternative is in the form H_A: [Benzene] ~ Normal(μ, σ^2), $\mu > 0$. Statisticians know this, but the UG doesn't quite manage to say it correctly. The potential for confusion is compounded on the next page by referring to the "normal distribution" [at 4-3], as if there were only one. I realize this abuse of language (speaking of a family of distributions in the singular) is commonplace in statistical writing, but it seems to be an ongoing source of confusion among the uninitiated or unwary. This is one place where we should take more care to be precise with the language.

Continuing with this discussion, the next paragraph asserts "In most cases, assigning to the observed data a low probability of occurrence under H_0 is cause for rejecting the null hypothesis in favor of H_A." This time, I think this won't cause any heartache among statistical novices, but it definitely should give statisticians problems. The discussion seems (implicitly) to be describing likelihood ratio tests. The critical point to make is not what the probability under H_0 is, but what the *relative* probability *densities* are under H_0

and H_A. It is unfortunate but demonstrably true that if the probability under H_A is sufficiently small, then the null hypothesis can be accepted even when it predicts a very low probability of occurrence as well. It's misleading to bury this key point (one of the bases for many criticisms of hypothesis testing) with the caveat "in most cases."

The UG does not correctly define the "false positive rate" [at 4-6]. This is a crucial concept that many readers will misunderstand. Part of the problem is that at least three distinct things could be described by "false positive rate:" (i) the *expected* rate of false positive results, given the true underlying distribution; (ii) the *maximum* expected rate of false positive results that could be attained by some element of the null hypothesis[9]; and (iii) the *observed* rate of positive results that are subsequently shown not to indicate contamination (this is what many readers will mistakenly understand it to be). I recommend that the discussion clarify these distinctions and provide the correct definition (which is (ii)). In particular, in most applications, "a test run at the $\alpha = 0.01$ level of significance" means that prior to gathering the data, there is *at most* "a 1% … probability that a type I error will occur in the results" [at 4-6].

Example 4-1 [at 4-7] attempts to clarify the situation, but it does not help much, because it could too easily be misinterpreted. It refers to "values of the chi-square test statistic" but these are confusingly labeled in Figure 4-2 [at 4-8] as "concentration." We are told that these values "become less and less probable as they increase in magnitude." Won't many readers then assume that the height of the curve (the pdf) gives the probability, especially because the UG has not defined what distributions are or how to understand a probability distribution function? From the text it is nowhere clear that α refers to a tail area rather than the height of the curve. The Figure does not resolve this ambiguity.

It is nice to see a discussion of the Central Limit Theorem (CLT) in the introduction. It's worth driving home the point that many tests will have asymptotic Normal-theory versions that allow one to look up an approximate P-value in a standard table of the Normal distribution. This idea is well-conceived. Now, one doesn't want a guidance document to be too technical—especially in its introductory section—but the subsequent statement of the CLT unfortunately is too sloppy. At the very least it should qualify the "random variables" involved as being mutually independent. Even better, the statement should limit itself to identically distributed variables: that's good enough for the intended application. Too many people automatically assume that just about any combination of any random variables is approximately normal (a conclusion that often is false) and cite, in support of this, a version of the CLT remarkably like the statement in the UG. We need to be careful.

Sections II through IV

Please see the detailed comments in response to question 3 above.

[9] Much later, Section IV makes this distinction.

Are you aware of any other significant methods, approaches or strategies that are relevant and should be included in the document? Please explain, offer suggestions regarding where and how the methods/approaches/strategies could be incorporated, and provide relevant citations.

In addition to methods, approaches, and strategies described in responses to the preceding questions, I offer the following suggestions.

Chapter 7

The sole remedy suggested for unequal variances is "the data should be logged and retested" [at 7-9]. This is not a sufficiently general response. It could help to describe a simple diagnostic test, such as Tukey's spread-versus-level plot[10]. Plots often provide more precise, detailed information about the sources of problems and possible remedies. For example, a spread-versus level test would justify taking logarithms in Example 7-2, but it would also make clear that the need for logs depends critically on the values in Well 6 alone: the data in the first five wells do not suggest any transformation is needed.

[10] Essentially, fit a robust line to log(IQR) versus log(median), let its slope be *p* (usually rounded to the nearest half-integer) and attempt the Box-Cox (power) transformation of exponent 1-*p*. See John Tukey, *Exploratory Data Analysis*, 1977. (Chapter 4 or 5 as I recall.)

Specific Topics

1. The Unified Guidance presents a comprehensive approach to address the multiple comparisons problem in detection monitoring. Both sitewide cumulative false positive and false negative (power) errors are addressed, primarily through the use of prediction limit retesting strategies. Is this approach reasonable and sound? Please explain and offer any suggestions, as appropriate.

Overall this approach is well reasoned and, in many situations, will be sound. I raise some possible objections and exceptions elsewhere in this report, such as comments (5) and (7) of the Introduction (above) and my response to General Topics question 1a and to question 2 immediately below.

It appears that the UG makes a lot of progress in describing statistical procedures that better reflect monitoring realities and the nature of groundwater, but it still has not created a framework that meaningfully and fairly balances false positive and false negative rates. The UG does a good job in recognizing that error rates are really surrogates for *losses*: a material loss experienced by the facility when false positives occur and a loss of environmental protection experienced by all when false negatives occur. However, it only partially takes the next step, that of incorporating time calculations into its framework. A false positive leads to an immediate loss, whereas a false negative only delays, usually for a short period of time, the moment at which a release is correctly detected. That delay can translate to increased risk of environmental harm at some facilities while at other facilities it is of relatively little importance. Therefore, a truly comprehensive and fair framework must make a better accounting of the importance of sampling frequency and time for response. I hope that the promised research in control chart methodologies will begin to highlight these issues and provide better methods to address them.

2. The Unified Guidance concludes that similar cumulative false positive errors cannot be realistically defined for compliance/corrective action testing against a fixed standard. Two major recommendations are provided: 1) *a priori* power criteria to allow for consistent ability to detect increases above a standard under conditions of fixed or small sample sizes, and 2) aggregation of annual data to enhance both power and single-test false positive errors. For corrective action testing, enhancing power is left to the discretion of the facility beyond aggregating annual data, and a predetermined single-test false positive is recommended. Is this approach reasonable and sound? Please explain and offer any suggestions as appropriate.

There are several important issues to consider. One is the fact that monitoring occurs on a regular, predetermined, ongoing basis, so that the power of a single test underestimates the ability of the program ultimately to determine that concentrations exceed the GWPS. If the sequence of true mean concentrations over time is $(\delta_1, \delta_2, \ldots, \delta_n)$, if independent data are collected at each time, and if those data are tested with a power of $1-\beta(\delta_1), \ldots, 1-\beta(\delta_n)$, respectively, then the chance that *none* of them will trigger a statistically significant increase after *n* periods evidently is $\beta(\delta_1)\ldots\beta(\delta_n)$, which rapidly grows very small when all the deltas remain above zero. Furthermore, in situations where multiple wells are engaged in compliance monitoring[11] and any possible corrective action would likely affect them all, then the power of the monitoring program to lead to corrective action similarly increases as a function of the number of wells and the power to detect an elevated mean at each one of them.

This suggests that a full consideration of the capabilities of a RCRA compliance monitoring program should include a *time frame during which true exceedances need to be reliably detected*. Such a time frame would depend on the rates of groundwater flow, retardation factors of contaminants, proximity of compliance wells to downgradient receptors, and the potential ability of the permittee ultimately to perform an effective and timely corrective action should that become necessary. Because these are site specific and vary so much among facilities, it would not be appropriate for the UG to recommend particular values.

Another consideration is that the UG "does not recommend retesting in compliance monitoring" [at 16-33]. The reason it gives (namely, to argue that retesting is somehow built in to any evaluation of a mean or median) is not convincing: during compliance monitoring, changes often are occurring over time, so that the collection of observations often is *not* statistically independent. Furthermore, it frequently is the case that *all* observations collected during one monitoring period are *dependent*, because of variation caused by sampling and analytical procedures (a sampler and a laboratory effect, respectively). For this reason, the current practice of some facilities is to send samples for retesting either to a different laboratory or to two laboratories (the original and

[11] Similar reasoning applies, *mutatis mutandis*, to detection monitoring. It is modified by the consideration that mere *detection* of a release is not the same as *identification of concentrations above a GWPS*, suggesting that in general a facility's response need not be as rapid or reliable as it must be in compliance monitoring.

another one) in order to evaluate the laboratory effect, which can be large, especially for organics. It would be good for the EPA to recommend procedures which allow for this kind of option even during compliance monitoring.

The *a priori* power criteria for compliance monitoring recommended in the UG are so inflexible that in some cases they will be too stringent and in others not stringent enough. By proposing them, the EPA seems to be trying to force facilities to engage in much more frequent sampling and analysis during compliance monitoring than was envisioned during the passage of the RCRA regulations, in order to maintain an acceptably low false positive rate. However, in some circumstances, such power criteria would not be stringent enough. Suppose, for example, that compliance monitoring is being conducted against risk-based ACLs. Suppose further that those ACLs correspond to a 10^{-4} risk estimate and were approved by a regulatory body that desired to keep risk within a 10^{-6} to 10^{-4} range. Because the ACLs are at the upper end of the range, it could reasonably be argued that there must be effective power to detect *any* increase of the mean over the ACL, no matter how small, provided that the false positive rate when the true concentration is at the *lower* end of the range is sufficiently low. On the other hand, if the ACLs were established based on, say, a 10^{-5} risk on the grounds that it falls "in the middle" of a desired 10^{-6} to 10^{-4} range, then it could reasonably be argued that the monitoring program need only have high power to detect a mean that exceeds 10^{-4} / 10^{-5} = 10 times the ACL[12]

In short, the power we should be considering is not just the power to detect an elevated mean at one monitoring period at one well, but rather the power to cause the facility to move into corrective action during a pre-specified length of time. In most cases this power will be much greater than the nominal powers (of 50% or 80%) recommended in the UG. It is becoming clear that some form of sequential sampling theory (or control chart methodology) may ultimately deliver the best performance for many RCRA programs; in such a context, considerations of time frames (expressed in terms of the distribution of in-control and out-of-control run lengths) clearly are of paramount importance.

I am also not completely convinced that there is nothing to be done about estimating the false positive rate during compliance monitoring. The UG makes a good case—its reasoning is valid and important—but in my experience it is still possible to estimate a false positive rate. Usually there is at most a handful of analytical parameters that have any risk of being close to their GWPSs, as the UG asserts. Groups of these will have very strong correlations among each other. One can either look at each group as if it

[12] Indeed, it would seem that a sophisticated compliance monitoring program should specify not one GWPS, but two: a lower and an upper. When the true mean at a well is below the lower, the false positive rate should be very low; when the true mean is above the upper, the power should be very high. We run into problems because only a single GWPS is traditionally specified.

generated one independent datum during each monitoring round, or one can compute correlation coefficients based on recent data and make appropriate adjustments to the degree of freedom. Either way, it is possible to develop an "effective" number of monitoring parameters that should be treated as potentially having means just below their GWPSs during the lifetime of the permit, and then to use this number (which is often between one and five) to estimate a facility-wide false positive rate. My concern is that by summarily ruling out such approaches, the UG will make it much more difficult for them to be developed, proposed, or implemented.

It might be better were the UG to retain its discussion of the statistical issues and its expression of EPA's goals for compliance monitoring, *without* making specific recommendations that might narrow the range of options for innovative approaches in monitoring or statistical analysis. In particular, specifying 50% power at 1.5 "relative risk" and 80% power at 2.0 relative risk comes across as arbitrary and perhaps counter-productive. It would be useful to recommend that some such criteria be used during permit development, but without specifying exactly what the powers and the relative risks ought to be in every case. If numerical criteria are desired, then specify combinations of power and relative risk (such as 50% power at 2.0 relative risk) that are *feasible* (that is, can be achieved with low false positive rates) *using current practices*; namely, with monitoring that is no more frequent than quarterly (with the possibility of more frequent monitoring when obtaining verification samples). The nomograph [at 16-15] makes it clear that false positive rates no greater than 5% are achievable with sample sizes as low as 4 when 50% at 2.0 relative risk is desired (assuming Normal distributions).

3. Please identify any other recommendations that represent a revision and/or enhancement to current guidance and practice and that have the potential to significantly affect groundwater monitoring under RCRA or other environmental programs. For each topic identified, please answer the following questions:

a. Are the recommendations appropriate and reasonable given available methods, documented experience, and current practice? Please explain

b. Does the document provide adequate guidance to help owners and operators, Regional and State regulators, and others put these recommendations into practice? Please explain and offer suggestions, as appropriate.

Some of the previous responses address these questions. The following sections add to those responses by identifying revisions or enhancements that I have not yet specifically discussed.

Default assumption of Normality

The arguments for making Normality a default assumption (Section 6.2) are cogent and well thought out. I am not convinced, though, that the recommended implementation of this assumption is the best.

As far as I can gather—the UG is not particularly clear—the intention is to test all compliance and background data for Normality *every time any test will be conducted*. This contrasts with previous practice, which was to perform a comprehensive evaluation of historical data, especially data at upgradient wells, during the permit writing process[13]. This evaluation would produce conclusions, if only tentative, concerning the likely statistical distribution of compliance data (in the absence of overt contamination). These conclusions would influence the selection of routine tests and would govern assumptions about underlying distributions. Thus, for instance, if upgradient and historical data suggested Lognormal distributions, then one would adopt Lognormality as the default assumption. The UG appears not to allow for this reasonable, data-driven approach. Instead it seems to recommend assuming Normality of future compliance values *regardless of what was learned from past monitoring events*.

The default assumption of Normality carries over into the UG's computation of power curves, especially those for nonparametric prediction limits. Thus, it has a pervasive effect on both the practice of statistical testing as well as on the actual criteria the EPA uses to evaluate the quality of those tests. In situations where historical data indicate underlying distributions are unlikely to be Normal, it would be more reasonable (for the purpose of computing power) to adopt distributional assumptions consistent with the data rather than with the UG's proposed default.

[13] Some permits call for periodic reviews, often every two years, during which that evaluation would be repeated and default assumptions about statistical distributions might thereby be modified.

The UG provides useful methods for readers to evaluate the Normality of data, such as the Shapiro-Wilks test. They are some of the best. They are not the only ones, however, and they are limited to datasets of 100 values or less. One requires extensive tables of coefficients to carry out the Shapiro-Wilks test with fewer than 50 values, making its automatic application (in a spreadsheet or database environment) difficult. M. A. Stephens (JASA v.69 # 347 pp 730-737: *EDF Statistics for Goodness of Fit and Some Comparisons*, 1974) describes some approximate versions of Normality tests, such as the Shapiro-Francia, Anderson-Darling, and others, that are much easier to carry out and work almost as well (and sometimes better). It would be useful for the UG to acknowledge that these tests exist and to allow that they also have merit in groundwater monitoring applications.

Allowance for using power transformations

In many cases the UG acknowledges that data will be demonstrably neither Normal nor Lognormal. It suggests power (or Box-Cox) transformations in such cases. This is new and might be effective. (Certain groundwater analytes, such as fluoride, frequently appear to be square-root-Normally distributed.) The UG does not provide guidance for carrying out such an analysis. This choice is perfectly appropriate, given the already large size and scope: it would not be possible to detail every procedure that one might perform. But if one is going to entertain such a variety of transformations, it's a very small step to considering other distributional models, such as the family of Gamma distributions, or extreme-value distributions, and many others.

This possibility, if not controlled, opens the framework of the UG up to a particularly insidious abuse. It is not difficult to envision a reader, either uninformed or unscrupulous, routinely using readily-available software to fit any one of hundreds of statistical models to every set of data, automatically. Imagine an annual groundwater report in which various analytical parameters at various wells are characterized by a whole zoo of fitted distributions. For this reason, the latitude to fit a wide variety of distributional models to the data should be carefully controlled, limited perhaps to the permit-writing phase. For routine statistical testing, it would be best for the number of diagnostic tests performed (of distribution, homoscedasticity, percentage of nondetects, correlation, seasonality, and so on) to be severely limited. Otherwise computing actual false positive and false negative rates will become a nightmare. Presently, the UG is not clear about this. Many of its detailed "procedures," which appear to be descriptions of how tests should routinely be conducted, explicitly require such diagnostic tests to be performed.

Requirement for explicit power calculations

From the point of view of meeting RCRA objectives, this is an excellent requirement. If a facility is going to propose a statistical monitoring program, it should be able to demonstrate to the EPA's satisfaction that its effective power to detect an important release of contamination at any compliance well within a desired period of time is adequate. In many cases this will require sophisticated numerical integration or Monte-

Carlo simulation. But how many "owners and operators, Regional and State regulators, and others" will be able to put this into practice? After all, some 15 years ago Monte-Carlo techniques began to be popularized among risk assessors, a group that if anything has a better background to carry out such calculations than most owners, operators, and regulators. Unfortunately, many in the risk assessment community are still struggling with the most basic aspects of Monte-Carlo simulation, such as determining how many iterations are needed, modeling correlations, and performing basic checks to find calculation errors. This suggests that most readers of the UG will not be in a position to carry out most power calculations, at least not for a long time, *especially because the UG provides no guidance for doing so*. Attractive and reasonable as this recommendation to compute power is, its ultimate result may be to force most RCRA facilities to hire statistical consultants (or at least to buy their high-priced software). In this regard, offering software like the program for Chapter 13 is a very good idea. Perhaps the EPA should consider publishing additional software aimed at performing most of the power calculations that would be needed by any RCRA facility. If that is the intent, then consider funding a directed development effort so that ease of use and accuracy can be assured.

4. Are the statistical method summaries and flowcharts in Chapter 5 useful, and do they provide clear guidance for potential users?

The statistical method summaries are a welcome surprise: something like this has not appeared in previous statistical guidance, but it looks like a really good idea.

A closer integration of the method summaries with the flowcharts (through mutual cross-referencing) might enhance them both.

I have some minor editorial comments about the summaries (below). The most substantial is that as one reads through the summaries, the distinction between "hypothesis tested" and "underlying assumptions" seems to change. Early on, the "underlying assumptions" sometimes are equated with the null hypothesis, whereas later they genuinely are statistical assumptions. I also take "underlying assumptions" to include (beyond the theoretical statistical assumptions) certain properties of the data, such as allowable proportions of nondetects, that are needed for accurate application of each method.

Summaries

- ***Skewness coefficient*** [at 5-14]: The underlying assumption should include that there are no ND results. (Higher moments will become ever more sensitive to methods of imputing values to NDs.)

- ***Skewness coefficient*** [at 5-14]: Step (2) of "steps involved" should explicitly say that the *absolute value* of the estimated skewness should be compared to 1.

- ***Skewness coefficient*** [at 5-14]: Include a "when to use" section.

- ***Probability plot*** [at 5-15, top]: There are no underlying assumptions, apart from the ability to rank the data. This method is applicable regardless of how the data originated.

- ***Shapiro-Wilks test*** [at 5-15]: The underlying assumption is that there are no NDs in the dataset. (This test is going to be sensitive to ND imputation.)

- ***Filliben's test*** [at 5-16]: The underlying assumption is that there are no NDs in the dataset.

- ***Filliben's test*** [at 5-16]: The phrase "calculate the correlation between the pairs on the probability plot" is potentially confusing. (Pairs of points?) How about "calculate the correlation coefficient of the probability plot."

- ***Shapiro-Wilks*** multiple group test [at 5-17]: There are no underlying assumptions, apart from a lack of NDs in the dataset.

- ***Box plot*** [at 5-18, top]: The underlying assumption is that the middle 50% of the values are quantified; equivalently, there are fewer than 25% NDs in each dataset.

- ***Box plot*** [at 5-18]: Replace "at distinct well locations" in the "When to use" section by "among the different datasets." This would make it consistent with the "hypothesis tested" section and keep the description more general.

- ***Dixon's test*** [at 5-19]: The discussion in "when to use" implicitly, but strongly, implies that any outliers identified will be removed from the dataset. Rewrite this section to clarify that an outlier test *only* highlights a value for further examination, but *never* should be used for automatic deletion of data. (This is an important point: I have seen more than one consultant attempt wholesale automatic deletion of data, such as automatically removing anything further than two SDs from the mean of every batch! Let's anticipate that some readers may be similarly misguided.)

- ***One-way ANOVA*** [at 5-21]: The underlying assumptions also include that there be very few NDs and that the error terms be independent and identically distributed.

- ***Time series plots*** [at 5-23]: Please provide a chapter reference concerning how "adjustments can be made to the estimated standard deviation" (in the "advantages/disadvantages" section).

- ***Simple substitution*** [at 5-27]: The underlying assumptions seem too restrictive. In particular, it is not necessary to assume the median of the censored results is at one-half the QL. The real assumption is that any bias or change in test size or power experienced in follow-on procedures (such as computing prediction or confidence limits) that is introduced by the substitution method will not make a material difference.

- ***Simple substitution*** [at 5-27]: The "when to use" section seems to command the reader always to use simple substitution whenever a dataset contains fewer than 20% NDs. Replace "should be used when" by "should be used only when."

- ***Cohen's adjustment*** [at 5-30]: Could you say something about how many degrees of freedom to use when estimating the SD using this method? It is not clear that it should equal n-1, but at the same time simulations suggest that it is usually greater than the number of quantified values.

- ***Wilcoxon Rank-Sum test*** [at 5-33]: It could help to be more specific about what "same distributional form and … variances equal" means (here and in subsequent descriptions of nonparametric tests). It's unclear what

"distributional form" is intended to mean. For instance, do all Gamma distributions have the same "form" or not? Almost all the nonparametric tests assume distributions vary only in location, not in any other way. This remark therefore applies to the non-parametric prediction limit for *k* future values [at 5-36] and the Prediction limit for future median [at 5-40].

- ***Prediction limit for k future values*** [at 5-35]: The recommendation about selecting the value of *k* is confusing because it appears to define *k* in terms of itself. Please rewrite this. The same comment applies to the non-parametric prediction limit for *k* future values [at 5-36, bottom].

- ***Prediction limit for k future values*** [at 5-35]: Could you provide a chapter reference for Dunnett's MCC procedure? This comment applies wherever the MCC procedure is referenced in this section.

- ***Non-parametric prediction limit for k future values*** [sic; at 5-36]: Isn't this really a prediction limit for the *smallest* of *k* future values?

- ***Poisson prediction limit*** [at 5-42]: This test appears particularly susceptible to misuse and abuse. Stating that "the data must be re-scaled" moves in the right direction (towards warning the reader), but—by being vague and non-quantitative—does not go far enough. I recommend that (a) the reader be alerted that this test alone, of all tests described in the UG, depends critically on the units of measurement; and (b) that methods be described for testing whether the underlying distribution is Poisson. Consider recommending the "Poissonness Plot" described by Hoaglin and Tukey in *Exploring Data Tables, Trends, and Shapes* [Wiley, 1985], Chapter 9. Better yet, I think you should recommend *against* using this test, mainly because there is no legitimate scientific justification for its use in this context. (Concentrations reported by a laboratory are *not* counted data; even in the extremely rare cases where they might be counted data, the procedure of replacing NDs by one-half the DL probably destroys any Poisson character in the underlying distribution.)

- ***Sen's non-parametric estimate of slope*** [at 5-46]: The underlying assumption, according to Sen's 1968 JASA paper (equation 1.1), is that the distributions of the data are *identical* apart from a change in location that is a linear function of time. This assumption can be evaluated by examining the behavior of residuals over time.

- ***Sen's non-parametric estimate of slope*** [at 5-46 and again at 14-31]: The method is nonparametric not because the median slope is utilized, but due to the explicitly nonparametric distributional assumption (*q.v.*). (After all, utilizing the median is not confined to non-parametric estimation. In any parametric setting where all distributions are symmetric, for instance, the median is a robust, unbiased estimator of the mean. This illustrates how

"non-parametric" refers to the assumptions made about the underlying distribution, *not* about the form of the estimator.)

- ***At multiple locations***, beginning at 5-46, the word "pinpointing" is used to describe confidence intervals. This is potentially confusing because "pinpointing" seems to be the opposite of what a confidence interval does. Most writers use a word like "bracketing" instead. That would be a good choice here.

- ***Confidence interval on lognormal geometric mean*** [at 5-48]: The null hypothesis is that the true *geometric* mean concentration does not exceed the compliance limit. Similarly, the alternative hypothesis concerns the *geometric* mean.

- ***Confidence interval on lognormal arithmetic mean*** [at 5-50]: I am concerned that the discussion of "when to use" implicitly establishes a flexibility that ought not to be available. Specifically, this discussion seems to allow that the reader should feel free to change the very nature of a test when it suits her. If one needs a confidence interval around an arithmetic mean, but doesn't like one method of computing it, well then (the UG seems to say), change the *intent* of the test by computing something different: a CI around the geometric mean, say, or maybe a CI about the median. If you can do that, what's to stop you from going further and deciding you don't like any of those parameters and instead want to compute a CI around, say, the tenth percentile or something else that delivers convenient results? Better, wherever test assumptions might be violated, the UG ought to recommend different procedures for testing the *same* parameter, rather than recommend changing the test altogether[14].

- ***Confidence limit on upper percentile*** [at 5-53]: The final statement, *viz.*, "the user should always try to assess what a given standard represents before choosing which type of confidence interval to construct," appears consistent with the concerns I expressed above. This statement is so important that it deserves to be discussed at length earlier in the UG, perhaps in Chapter 4 ("Statistical Background"), rather than buried here.

- ***Confidence interval around trend line*** [at 5-55, bottom]: Practical considerations suggest this method should not be recommended for Corrective Action monitoring, at least not without modification. During CA, concentrations begin (usually) by exceeding the cleanup standard. After CA activities cease, the concentrations may continue to drop, may increase to a plateau ("rebound"), or stay essentially constant. (Other patterns are possible, but these three are common.) In any case, by including data prior

[14] Specific sections in later chapters do discuss most of these issues adequately. That discussion does not seem to be reflected in these summaries, though.

to the cessation of activities, it is virtually certain that the UCL of *some* sampling event will exceed the standard (if only at a distant time in the past). This problem is easily correctable, but the recommendations here do not acknowledge the problem or provide a remedy. The simplest remedy is to compute the UCL only for the most recent sampling event, but it's also important to allow the permittee to establish a time during which data will be collected to demonstrate attainment of the standard, rather than suggesting that all data be used.

Flowcharts

The idea of the flowcharts is a good one. Many readers will find these helpful. Analyzing a complex process into a couple dozen relatively simple charts helps make it more accessible; that is well done. The examples in this draft of the UG provide evidence of the potential efficacy of this approach.

I have a couple of general concerns about the use of flowcharts here, though, regardless of their clarity or quality. The first is that a flowchart could be interpreted rather strictly by many users, who would demand that any proposal for a statistical test conform to the flowcharts. I appreciate that the footnote [at 5-1] provides a "strong recommendation" to refer to additional material, but believe this does not go far enough. Embedding it in a footnote reduces the strength of the recommendation. The recommendation, as it stands, is somewhat elliptical (and even patronizing), stating that the "detailed textual material" should be "thoroughly digested." (Are we pushing the cookbook analogy a little too far?) It would be better to see a strong caveat appear in the text itself, and perhaps in a caption to each flowchart, stating that the flowchart is provided only as general guidance, is by no means complete, and does not constitute an EPA requirement.

The second general concern is that because new material seems to appear in the flowcharts, especially in the decision boxes, which appears not to be explained elsewhere, it could be misunderstood and misused. For instance, what is the basis for the various percentage-of-nondetect cutoffs that appear? Whence come the 20%, 50%, and 75% in Flowchart 5-18 (to take one of many examples)? Exactly how should these percentages be computed? Are they to be taken as hard-and-fast rules or as approximate guidance, to be modified by statistical judgment as circumstances dictate? If answers to these questions do appear elsewhere in the Guidance, then please provide cross-references within the decision boxes themselves.

I do have some specific comments and questions about the flowcharts in their current form.

What is the meaning of the different shapes used in the flow charts? Trapezoids, parallelograms, rectangles, and other symbols (which have traditional meanings that are inappropriate or misleading in this application) abound, but I can discern no functional distinction among them. Rectangles appear to have multiple functions, including procedures and connectors to other charts. Some symbols have dark backgrounds (as in

Flowchart 5-6, for instance). All this is confusing. It appears these flow charts have four elements: termini (start/stop), tests (diamond shapes), procedures (all other shapes), and connectors (rectangles). Adopt a conventional association between shapes and functional elements. Use this consistently.

The level of aggregation or abstraction in the flowcharts varies erratically and inconsistently. In some cases an instruction amounts to a substantial project in itself, such as "[determine whether] intrawell tests [are] adequately powerful" (Flowchart 5-3), whereas in other cases the instruction is a trivial computation executed as a part of a test, such as "sort compliance data and compute ranks" (Flowchart 5-8, *e.g.*). Make the flowcharts more uniform in this regard.

The flow charts are cryptic and incomplete. Fix these problems through revision, checking of cross-references, and thorough editorial review. Let me provide just a few examples from the very many that could be produced:

- What exactly is the distinction between a "Risk-Based MCL" and a "Mean-Based MCL" in Flowchart 5-1?

- When one evaluates an "apparent trend in compliance data" [in Flowchart 5-1, two places], exactly which data should be used? Normally, after corrective action one does not use all data, but instead attempts to identify a point at which conditions have reached a stable, quasi-steady state; previous data are not used for trend analysis. At what significance level should such a trend be tested? Should the significance be adjusted for multiple comparisons?

- When one "compute[s] overall ND% across all wells" in Flowchart 5-2, is this done parameter by parameter or not? Are only "background" or "baseline" data used, or are all data used? Is a detected below-PQL value considered an "ND" or not? What if some wells have samples that are always ND and all other wells consistently have detectable results? Would the answer depend on whether the ND well is a background well or a compliance well?

- What is the basis for the 20% criterion for ND% in Flowchart 5-2?

- In Flowchart 5-2, what does the "spatial variation uncertain" procedure mean? Is it telling us to do something?

- How does one reconcile the apparent contradiction in the left path of Flowchart 5-2, where first we are told "no statistical test needed" and later we are told "SSI [statistically significant increase] = 2 consecutive quantified detections?" How can one have an SSI without a statistical test?

- What exactly is the mechanism for moving from one flowchart to another? For instance, in Flowchart 5-1, we are exhorted to "check intrawell vs. interwell Flowchart 5-3." Moving to the start of Flowchart 5-3, there is a path terminating at "tests for trend Flowchart 5-7." All paths terminate at "Sen's Slope estimate Chapter 14." The problem is not yet solved, though. Should we backtrack? This returns us to Flowchart 5-1, but the very next step is "Upgradient to DG comparison OK?" We have not yet been shown how to select a test for making this comparison. We're stuck.

- On Flowchart 5-5, what is "Sort Interwell BG" supposed to mean? What does one do?

- On various Flowcharts, beginning with 5-5, there are lettered circles "A", "B", "C", and "D." Some of these are termini; some are not. By convention such symbols are used to jump to other locations within a chart or among charts, but evidently that makes no sense here: for instance, there are two unlabeled branches emanating from the "D" in Flowchart 5-18. What is intended by these symbols?

- The branches on Flowchart 5-7 are missing their labels.

- In Flowchart 5-7, does "NDs present" mean "one or more NDs"? This seems unnecessarily restrictive.

- Flowchart 5-6 refers to nonexistent Flowchart 5-20 (variously called "Intrawell Control Chart" and "Intrawell control chart on transformed values").

- There is no flowchart that refers to Flowchart 5-18, "Prediction Limit for Mean/Median (w/ Retesting) Part 1."

- In Flowchart 5-18, it is unclear which statistic is intended in phrases like "compliance well median" or "compliance well mean." Are these the median/mean of all compliance well data or of recent data from a single compliance well, for instance?

As these examples should make evident, the flowcharts do not provide clear guidance in their current form. To improve them, undertake a thorough revision for consistency, completeness, and clarity. Make sure that all procedure boxes explicitly reference part of the guidance (either a chapter or another flowchart). Provide a key to the flowchart symbols. Make the level of procedural detail consistent. Finally, reproduce the flowcharts in a more legible format (*e.g.*, direct printing to pdf or output via a metafile, rather than the low resolution bitmaps currently used.)

5. Is the software program for Chapter 13 non-parametric prediction limit testing useful and accurate?

This program's dialog is well laid out. It executes quickly. It is simple, focusing on one well-defined task. The underlying methodology for estimating power is clever and efficient. These are all commendable features.

I do, however, have problems using the program.

The dialog's prompts and the associated documentation are confusing. I am at a complete loss to understand what the *MX* term is: "the lowest maximum value of a data set which can be used for comparison." Huh? What "data set"? How can a dataset have more than one "maximum value," much less a lowest among them? "Comparison" to what? Is one supposed to input a concentration or a rank? Do the ranks go from lowest to highest or highest to lowest?

The description of the "calculate" function is even more baffling. What are "maximal values"? What does it mean for "absolute maximum = 1"? When the program outputs a message like "Alpha OK at maximum = 5," what exactly does this mean? Precisely what test is being suggested?

Through careful parsing of the documentation, I can glean answers to most of these questions. The point is that such hermeneutical sleuthing should not be necessary: make the dialog and the documentation clear, explicit, to the point, and readily understandable by the intended audience.

It is not possible to test the accuracy without understanding what algorithm is being used or what it is being used for. Neither the UG nor the program documentation describe the algorithm, except to remark that it is an approximate one: that just reinforces the need to document the algorithm with precision and accuracy.

I can guess at much of what this program is doing, but only through a substantial "reverse-engineering" process of typing in inputs, looking at the inputs, making educated guesses about what's going on, and then checking them against tables and my own calculations. Surely we cannot expect that of most users, except perhaps the most desperate among them. Because this is a program that will be used occasionally and whose results must be accurate, it should provide clear, unambiguous prompts, extensive feedback, and clear indications about how to interpret its results. In its present form, it supplies none of this.

The program happily accepts erroneous input. For instance, typing "10%" for the "single-constituent false positive" [sic] does not create the same output in general as typing "0.10". Fractional values for sample sizes are quietly accepted. Some values, such as negative background sample sizes, generate runtime errors and crash the program. Some typographical errors crash the program, but others, such as typing "1oo" instead of "100" for background sample size, actually produce output, which has to be

considered erroneous (there is no reliable way to know what input the user really intended). The program needs to trap all such input and respond informatively[15]. It would be best for the dialog to automatically re-display the input *as interpreted for the calculations* back to the user.

It is surprising that although the documentation can describe some invalid combinations of inputs, the program itself nevertheless contains no run-time checks or protection against such inputs: it just crashes. That's decidedly unfriendly. It also makes the program's output very suspicious: if the programmers are not trapping obvious errors, then perhaps they are not trapping less obvious errors, either, the ones that simply create invalid results. Because of this adverse behavior, this program simply cannot be trusted.

Note that the foregoing is criticism of the program interface, not the underlying algorithm. But without a reliable interface, one cannot rely on any of the outputs.

The program needs to be updated to incorporate the current guidance in the UG. In particular, the evaluation of power at 2 sigmas should be removed and the criteria of 0.55 and 0.82 and three and four sigmas should be changed to 0.5 and 0.8, respectively. Better yet, let the user input the sigmas and the power thresholds.

The program's documentation should describe exactly what algorithm is being used. Its description of the inputs and outputs should use the same terminology as the UG. It should include several worked examples, interpreted in plain language (or at least language consistent with the UG). For instance, there is nothing in the phrase "one could expect the … 20th highest maximum to be the optimal maximum for the higher level 1:m tests" [at page 5 of OptRankSummary.pdf] that makes any sense to me, so I expect many readers would have the same problem. One possible interpretation of this sentence might go like this: "In many situations where background sample sizes are large (above 100 or so), using the second-, third-, …, or possibly even the 20^{th}-highest background value as a prediction limit will produce the best balance of false positive rates and power for tests involving several resamples."

Finally, the dialog and its documentation should both make clear that the power calculations assume that all underlying distributions are Normal. This is often invalid in practice (or simply cannot reliably be checked due to high censoring levels).

My recommendations for improving this program so that users can obtain reliable results therefore are:

(1) Trap *any* non-numerical text that is entered in any of the dialog's input fields. Modify the dialog to consider any such text to be an input error and to issue an informative and helpful message.

[15] My testing suggests the program is extracting the largest prefix of each input string that can be interpreted as a number, after ignoring any blanks. Input entered in scientific notation, such as "1.0 e+1", has its exponent ignored.

(2) Enhance the dialog's input mechanisms so that only valid combinations of inputs are accepted and all other combinations produce an informative message that (a) does not crash the program and (b) suggests how to fix the problem.

(3) Augment the dialog's output so that it includes a clear summary of how the input fields were interpreted.

(4) Rewrite the dialog's prompts and labels so that they are clear, meaningful, and written in a terminology consistent with the UG.

(5) Add the ability to change the power criteria in the dialog: specifically, to modify the deltas (currently equal to two, three, and four) and the thresholds associated with them (currently 0.25, 0.55, and 0.82, respectively).

(6) Rewrite the supporting document (OptRankSummary.pdf) completely in clear, plain, conventional language, as in the example provided above.