APPENDIX C (DRAFT)

STATISTICAL CONSIDERATIONS FOR DATA QUALITY OBJECTIVES AND DATA QUALITY ASSESSMENTS IN WATER QUALITY ATTAINMENT STUDIES

Michael Riggs, Dept. Statistical Research, Research Triangle Institute

Acknowledgements:

We are grateful to Andy Clayton (RTI), Florence Fulk (EPA/NERL), Laura Gabanski (EPA/OWOW), Susan Holdsworth (EPA/OWOW), Forrest John (EPA/Region 6), and Roy Whitmore (RTI) for providing thoughtful reviews of earlier drafts of this appendix and for suggesting a number of changes which greatly improved the final version.

Appendix C Table of Contents

Appendix C Table of Contents (continued)

Appendix C
Statistical Considerations for Data Quality Objectives and Data Quality Assessments in Water Quality Attainment Studies

C.1 The Data Quality Objectives Process: Principles of Good Study Design

C.1.0 Introduction to DQO Procedure

The process of determining if a body of water meets water quality standards can be divided into two phases. The first phase, the study design phase, encompasses seven activities involved in specifying the appropriate research questions and developing a strategy for collecting the data needed to answer them. The seven activities comprise the Data Quality Objective (DQO) process. Part C.1 of this appendix details the four "statistical" DQOs (i.e., DQOs 4-7) and provides guidance for choosing the appropriate statistical tools for achieving them. The second phase, called the Data Quality Assessment (DQA), includes all of the statistical procedures necessary to answer the questions of interest in the face of the uncertainty inherent in the data and the data collection methods. Parts C.2 and C.3 of this appendix provide guidance for selecting and implementing inferential statistical techniques needed to complete a DQA to support a water quality attainment decision. Part C.2 introduces a collection of techniques called exploratory data analysis (EDA) that is useful for determining structure and pattern in the data and for examining the validity of various assumptions that may underlie the more formal processes of statistical inference (i.e., interval estimation and hypothesis testing) that are developed in Section C.3.

C.1.1 Review of the Basic DQO process

EPA defines environmental data to include data collected directly from field measurements or compiled from existing databases or literature. When such data are used to select between two alternative conditions (e.g., attainment or non-attainment), EPA requires that organizations responsible for monitoring ensure that the data used to characterize the environmental processes and conditions of concern are of the appropriate type and quality for their intended use and that environmental technologies are designed, constructed and operated according to defined expectations. The Data Quality Objective (DQO) process is EPA's recommended tool for developing sampling designs for data collection protocols that will ensure these conditions such that the quality of the data are sufficient to support decision-making within tolerable levels of decision error. The 7-step DQO process is described in detail in the EPA document, *Guidance for the Data Quality Objective Process* (EPA/600/R-96/055). The steps are:

1. Define the problem
2. Identify/state the decision which must be made
3. Identify the information (data) needed to make the decision
4. Define the target population
5. Develop a decision rule and the population parameters needed to make the decision
6. Specify the tolerable limits of error in making the decision
7. Choose an optimal sampling design

Step one involves the formulation of the general problem and a corresponding conceptual model of the hazard to be accessed. For example, we may want to examine a particular reservoir to estimate the mean selenite concentrations in the water column. Having identified the problem, it is then necessary to select a project manager, technical staff and the associated resources that will be needed to collect the appropriate data from the target waters.

In Step 2 of the DQO process, the water quality attainment (WQA) question and the alternative actions to be taken in response to the answer, must be clearly stated. The question will often be as simple as, “Does this reservoir attain the current WQA criterion for selenite concentration in the water column?” Alternative actions, may include listing the reservoir and/or instituting other remedial activities or use restrictions. The WQS question and the alternative action(s) together comprise the “decision statement” which is critical for defining decision performance criteria in Step 6 of the DQO process.

In Step 3, the specific field data that are needed to resolve the decision statement are identified, as well as any pertinent data in the literature or historical databases. At this time, the appropriate field and laboratory measurement and analytical methodologies should also be identified. The specifications and criteria for these measurements, measurement devices, laboratory procedures, etc. may be refined later in the DQO process.

Steps 1-3 require consultation among the planners and regulators and are primarily concerned with identifying qualitative criteria for the study. Once these have been established, the remaining 4 criteria must be addressed through consideration of statistical principles for population estimation.

C.1.2 Defining the Target Population

The target population is the entire set of elements to which the investigators desire to extrapolate the findings of their survey or monitoring design. Thus depending on the study objectives (DQOs 1 & 2), the target population may be all the waters in a state, a specific type or class of waters in a state, all the waters within a specific drainage, or only the waters in a particular stream reach or pond. Additional background on monitoring designs and target populations are presented in Chapter 12 of this document.

For environmental studies, it is critical that such populations are bounded in space and time. For example, investigators may desire to quantify the condition of a specific stream-segment in January of a specific year. Whereas in most biological studies the population elements, generally called sampling units, are discrete individuals (e.g., fish or people), in water quality studies (WQS), they are often volumes of water. Thus, if a **sampling unit** were defined as a 1-liter aliquot, the target population for a January, 2001 water quality assessment of a 3-mile reach of a stream would contain all of the 1-liter quanta, which flowed through the stream segment during January of 2001. Estimates from such a sample are only valid for inferences about that particular 3-mile reach in January 2001. They cannot provide a statistical basis for inferences to the same stream reach at any other time (e.g., August 1999), nor for other reaches of the same stream or for different streams.

C.1.3 Developing a Decision Rule and Choosing Parameters and their Sample Estimators

The population factor of interest is called a **parameter**; e.g., the proportion of sampling units in the target population which exceed some standard. The surest way to measure a population parameter is to make the measurement on every member of the population. The process of doing so is called a **census**. While this may be feasible for a small closed population (e.g., all the palm trees on a small island), it is not likely to be so for most bodies of water. Typically, some subset of the population will be selected as representative of the target population and the measurement will be made from this subset or **sample**. The measurement obtained from the sample (e.g., the sample proportion) is called a **statistic**; it provides the investigator with an estimate of the more difficult and expensive population parameter. To avoid confusion, in this appendix "sample" will be used in the statistical sense; i.e., it will refer to the collection of the individual volumes taken from the target water. The volumes themselves (e.g., 1-liter aliquot) will be referred to as "sampling units".

The choice of the criterion for whether a water body or stream segment has attained water quality standards or is impaired will generally be based on prior scientific investigations. Although such criteria exist for biological, physical, and chemical components of an aquatic system, this appendix focuses on the chemical components. For toxic (e.g., arsenic) pollutants, non-priority chemicals, and physical parameters (e.g., pH), EPA has established two types of criteria: acute and chronic. Acute criteria are based on 1-hour mean chemical concentrations determined from laboratory studies. The acute hourly means are called Criteria Maximum Concentrations (CMC) and have been tabulated for both toxic and non-priority pollutants (EPA 1999; http://www.epa.gov/OST/standards/wqcriteria.html). EPA regulations specify that the acute criteria for a pollutant must not be exceeded more than once in a given 3-year period for any water body. Chronic criteria are based on 4-day (i.e., 96-hour) chemical means, called Criterion Continuous Concentrations (CCC), obtained from laboratory and/or field studies. EPA regulations require that (1) the 30-day mean concentration of a pollutant in a body of water must not exceed the CCC for that pollutant more than once in a 3-year period and (2) no 4-day mean can exceed 2×CCC in any 4-week period. The limiting concentrations, durations and 3-year frequencies specified in the criteria are based on biological, ecological and toxicologic modeling studies and have been designed to protect aquatic organisms and ecosystems from unacceptable effects. Details regarding the calculation and scientific bases for CMC and CCC have been documented by Stephan et al. (1985).

Statistical methods must be employed to determine if the acute and/or chronic criteria for parameters such as dissolved oxygen, pH, surface temperature, etc. have been exceeded. For example, it may be known that concentrations of chemical X greater than 10 μg/l represent an important threshold for algal blooms. Given such a threshold, it is then necessary to define a sample statistic to which it may be compared. The choices may include: the maximum tolerable proportion of sampling units in the sample that exceed the threshold (e.g., 10%), the sample mean concentration, the sample median concentration or some percentile (e.g., the 95th) of the sampling unit concentrations in the sample. The choice of which statistic is "best" depends on the expected behavior of the sample statistics and on the study objectives. For example if the distribution of the concentrations among the sampling units is expected to be bell-shaped (or approximately so) then the mean may be the best choice. However, if the distribution is expected to be skewed (e.g., most sampling units have low concentrations but a few have very high concentrations), then the median might be preferred. On the other hand, the investigators may wish to make their decisions on the basis of extreme values, however rare in the sample, in which case the 95th percentile might be the appropriate statistic.

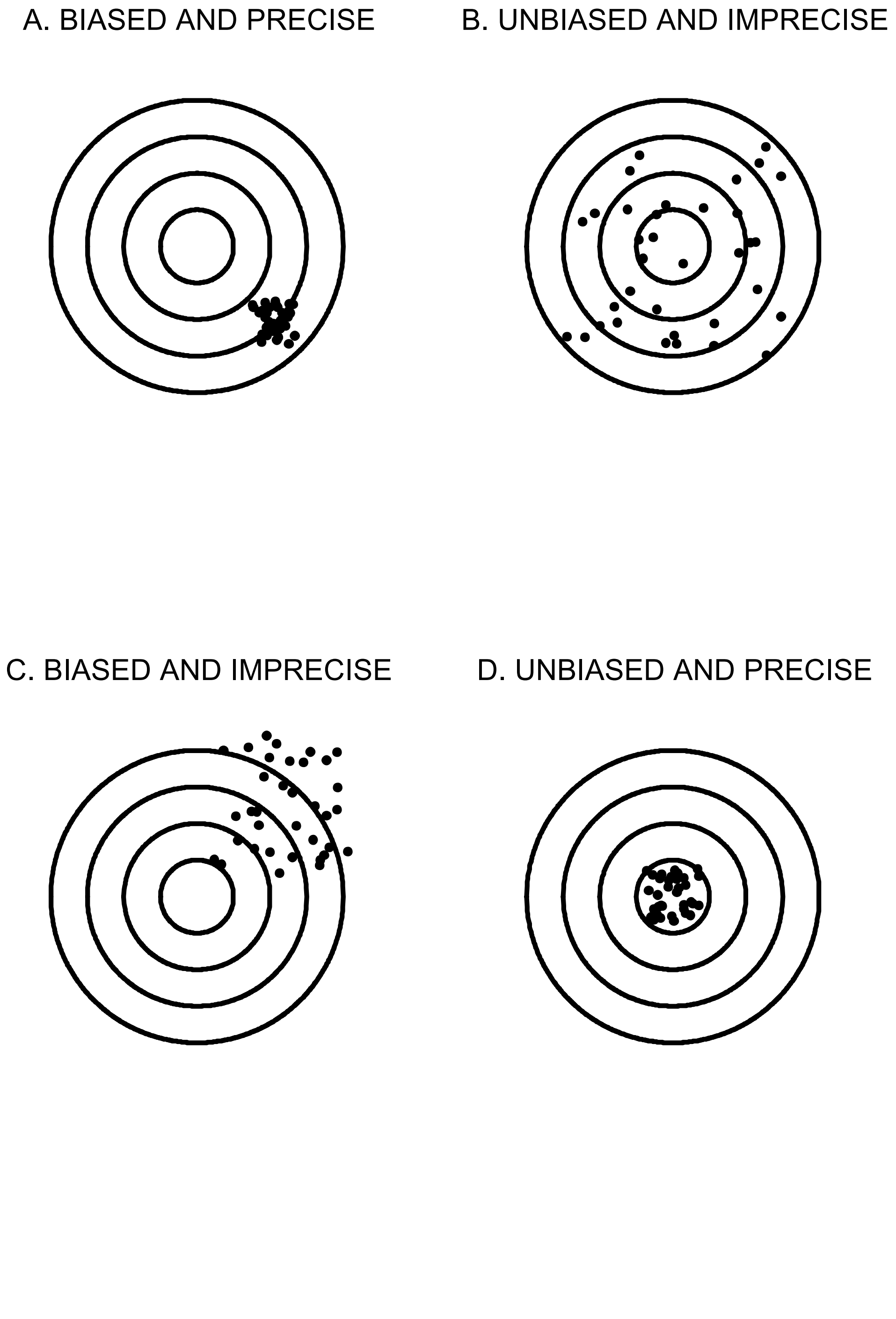
The choice of whether to measure central tendency or extreme values in a population distribution usually depends on whether we want some convenient means of characterizing the "average" condition in the population or if we want to know the magnitude of the "best" or "worst" conditions in the population. In the former situation we would focus on the mean or median. For example if some sort of remedial action had been applied to reduce nutrient loading in a body of water, we might want to compare mean or median concentrations of various phosphates or algal abundances before and after treatment to determine if there had been a "general" improvement in the body of water. Alternatively, if we had human health concerns associated with consumption of fish tissue containing mercury above some threshold concentration, we might want to estimate the 95th percentile of tissues concentrations of mercury in the resident fish population. The reason for this is that if 5% or more of the fish have tissue levels above a critical threshold for human health effects, there would be at least a 1/20 chance of toxic exposure due to human consumption of those fish. This would be so regardless of the magnitude of the population mean or median mercury concentration. In other words, interest focuses on the extremes (e.g., the 95th percentile) because it is only the extremes that are likely to effect human health.

C.1.4 Bias, Imprecision and Decision Error

Decisions in WQS will nearly always be based on sample data. Because sample data are subject to both bias and imprecision, decisions based on such data will be subject to error. To understand how to control this error, one first needs to understand the nature of bias and imprecision in sample-based inference.

It is intuitive that not every sample estimate is good, in the sense that its value is close to that of the unmeasured population parameter. In particular, the sample statistic may be biased, imprecise or both. Figure 1 illustrates four possibilities as distinct shot patterns around a bull's eye; each bull's eye represents a population target parameter, while the shots are statistical estimates of the parameter obtained from repeated sampling of the target population. Figure 1a illustrates the case where the statistical estimates are biased but precise. Sampling **bias** results from systematic error caused by sampling which favors some individuals over other population members. For example radio talk-show call-in surveys tend to over-represent disgruntled listeners. Thus their responses deviate from the true population parameter in the same direction (i.e., are biased) and tend to be more alike. This homogeneity is reflected by the tight clustering of the shots (i.e., the shot pattern is precise) in Fig. 1a. The statistical estimates in Fig. 1b are not skewed in any particular direction, thus there does not appear to be any bias. However the pattern is highly dispersed around the true parameter value. This is indicative of considerable heterogeneity in the target population, which leads to **imprecision** in the sample statistics. Imprecision and heterogeneity are reflected in increased dispersion of the statistical estimates

Fig. 1. Bias and precision as represented by shot patterns on a target. Each bulls-eye represents the true target population parameter and the shots represent sample estimates of the target parameter.



A. BIASED AND PRECISE

B. UNBIASED AND IMPRECISE

C. BIASED AND IMPRECISE

D. UNBIASED AND PRECISE

about the value of the target parameter. In situations like Fig. 1b, a large number of shots (i.e., samples) are needed to ascertain the approximate center (i.e., true parameter value). Whereas Fig. 1c depicts the worst case scenario (biased and imprecise), Fig. 1d illustrates the best case (unbiased and precise). Note that in Fig. 1d, a much smaller number of samples are needed to locate the bull's eye than in Fig. 1b; in fact, a parameter estimate based on any one of the samples in Fig. 1d will be correct, while one would need to average over all of the samples in 1b to get a parameter estimate that was close to the true target value.

Water quality attainment decisions are made on the basis of probabilities derived from the sample estimates by the application of statistical inferential procedures. Thus, bias or imprecision in the sample estimates may lead to erroneous probability statements. The computed probabilities support the acceptance or rejection of two competing hypotheses: the **null hypothesis** (H_0) that the water attains WQ standards vs. the **alternative hypothesis** (H_a) that the water is impaired. Given these two competing hypotheses, imprecision and/or bias in the sample estimates of the desired population parameters creates the potential for two types of decision errors.

So-called **Type I errors** occur when the null hypothesis is incorrectly rejected; i.e., a water that attains WQ standards is erroneously judged to be impaired. A **Type II error** occurs when an impaired water is erroneously judged to be in attainment. Type I and II errors are often compared to the judicial errors of convicting an innocent man (Type I) and letting a guilty man go free (Type II). Notice the near inevitability that when we decrease the probability of one error we coincidentally increase the probability of the other. Thus a subjective decision is usually made to guard against one at the expense of the other. For example, the U.S. judicial system has traditionally protected against convicting the innocent at the expense of letting the guilty go free. The Stalinist Soviet Union took the opposite view and jailed a large proportion of its population, innocent and guilty alike, with the result that nearly all criminals were sent to labor camps and Soviet crime rates were exceedingly low. Similar to the U.S. judicial system, the scientific community has focused on protecting against Type I errors rather than Type II errors. This is apparent in the published scientific literature wherein Type I error rates (the α-level) are almost always reported, while the Type II error rates (the β-level) are much less commonly considered.

The statement about the near inevitability of increasing β whenever α is reduced is conditional. To be strictly correct, we must say that *for a given quantity of evidence* (i.e., the sample size), decreasing α will inevitably increase β. This suggests that it is indeed possible to decrease α and β simultaneously, if one increases the amount of evidence. In the statistical assessment of water quality attainment issues, the sample is the ultimate source of the evidence. Thus, if we increase the number of sampling units (n) in the sample until it equals the population size (N), α and β will decline to zero. Unfortunately, sampling units are expensive to collect and process. Consequently the cost of simultaneous control of α and β to low levels (e.g., 0.05) is generally prohibitive, often requiring hundreds or even thousands of sampling units per sample.

C.1.5 Quantifying Sampling Error: Confidence Intervals

Clearly the Types I and II decision error rates reflect the imprecision in the sample estimates. The imprecision in the sample estimate, called **sampling error,** is due to approximating the true

population parameter (e.g., the mean) with an estimate computed from a sample. The **standard error** (SE) of a sample estimate of a population parameter provides a quantitative expression of the sampling error. For example, the standard error of the sample estimate $\left(\bar{x}\right)$ of the population mean (μ) is computed as shown in Box 1. Notice that as the sample size (n) increases to values which approach the population size (N), the correction factor (fpc) goes to zero and, further, when fpc=zero, the standard error will also be zero. This provides a mathematical justification for our earlier statement that the decision error rates, α and β, will be zero whenever the population is censused. Another way of saying this is that sampling error only exits when a population is sampled. As previously stated, in the context of water quality attainment studies, the sample size (n) will always be extremely small
relative to the population size (N). In such cases, the FPC is essentially equal to 1.0 and therefore is ignored. However this suggests that the standard errors of estimates obtained from water quality attainment sampling efforts are not trivial and hence the potential for substantial Type I and II decision errors must be addressed in the DQO process (step 5).

One way that statisticians deal with sampling error is to construct **confidence intervals** about the sample estimates such that the interval has some known probability (e.g., 95%) of containing the true population parameter. The confidence interval is a statement about the confidence we have in the sample estimate of a population parameter, θ. Algebraically, this statement is written as shown in the first expression in Box 2. By convention, the **confidence level** is expressed as a percent (e.g., 95%). For example if we desire to hold the Type I error rate to α=0.05, we are in effect saying that there is a 5% chance that our estimate is incorrect; thus, the corresponding confidence interval says that we can be 95% confident that the value of the unknown population parameter is within the bounds of the confidence interval. A more mathematically rigorous explanation of the 95% confidence interval is as follows: if one drew 100 different random samples from the population and computed 95% confidence interval estimates of the population parameter θ, from each sample, we would expect that 95 of the computed confidence intervals would include θ. Because this interpretation of the confidence interval is based on the concept of repeated sampling of the target population, this approach to statistical inference is called **frequentist statistical inference**.

Similar to hypothesis tests (see Section C.3.1), confidence intervals may be two-sided or one-sided (Box 2). The first expression in Box 2 is a two-sided 1-α confidence statement. Upper and lower 1-α one-sided confidence statements have the general form shown in the second and third expressions in Box 2. Two-sided confidence intervals are appropriate when one desires a sample estimate (with 100×(1-α)% confidence) of an unknown population parameter, as is the case in most investigations of processes that characterize natural populations and/or ecological systems. One-sided confidence intervals are appropriate when one wants to compare the sample estimates to a specific regulatory standard. For example, a lower one-sided confidence interval would be appropriate for comparing an observed chemical concentration to the maximum allowable concentration of that chemical. Conversely, an upper

Box 1-a: The Sample Mean and Its Standard Error

Suppose X represents the population factor of interest. Let *N* denote the population size, μ the population mean, and s^2 the population variance. Let X_1, X_2, …, X_n represent *n* data points, i.e. a random sample of *n* unit from the population of *N* units.

The sample estimate of the population mean μ is the sample mean $\overline{x}$:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} \ x_i \ .$$

The standard error $\overline{x}$ is given by:

$$SE(\overline{x}) = \sqrt{\frac{s^{\,2}}{n}\left(1-\frac{n}{N}\right)} = \frac{s}{\sqrt{n}} \ x \ \sqrt{1-\frac{n}{N}}$$

where s is called the population standard deviation and the quantity $1-\frac{n}{N}$ is the finite population correction factor (fpc). Oftentimes, the population variance s^2 is unknown but may be estimated using the sample variance s^2:

$$s^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{x}^2\right)}{n-1} \ .$$

Hence, the standard error of $\overline{x}$ may be estimated by replacing s^2 with s^2, or with s, the sample standard deviation.

Box 1-b: Example for Calculating the Sample Mean and its Standard Error

Consider a random sample of 10 of the 244 monthly turbidity measurements taken from the Mermentaut River between June 1980 and April 2000. The measurements (in NTU) were: 34, 58, 87, 145, 14, 38, 62, 95, 160, 320.

$$\bar{x} = \frac{34+58+87+145+14+38+62+95+16-+320}{10} = \frac{1013}{10} = 101.3$$

$$s^2 = \frac{(34-101.3)^2+(58-101.3)^2+(87-101.3)^2+\ldots+(160-101.3)^2+(320-101.3)^2}{10-1}$$

$$= \frac{4529.29+1874.89+204.49+\ldots+3445.69+47829.69}{9} = \frac{73006.10}{9}$$

$$= 8111.7889$$

$$s = \sqrt{8111.7889} = 90.07$$

Note that an infinite number of possible turbidity measurements could have been taken from throughout the target area during the 20-year monitoring period. In these circumstances the finite population correction factor is essentially 1.0 and can be ignored. Thus the estimate of the standard error is $\bar{x}$ is simply:

$$\hat{SE}(\bar{x}) = \sqrt{\frac{8111.7889}{10}} = 28.48\,.$$

Box 2: General Forms for 100 x (1-a)% Confidence Intervals

Let ? denote a population parameter. A two-sided 100x (1-a)% confidence interval about a sample estimate of ? has the general form:

$$\Pr\left[a_1 \le q \le a_2\right] = 1 - a$$

where 1-a = the desired confidence level
a_1 = the 1-a/2 lower bound of the sample estimate of ?
a_2 = the 1-a/2 upper bound of the sample estimate of ?.

One-sided confidence intervals have the general form

$$\Pr\left[-k \le q \le b_2\right] = 1 - a \quad \text{for an upper one-sided confidence interval}$$

$$\Pr\left[\mathrm{b}_1 \le q \le +k\right] = 1 - a \quad \text{for a lower one-sided confidence interval}$$

where -k = 0.0 for proportions and variances; -8 for means and medians
+k = 1.0 for proportions; +8 for means, medians and variances
b_1 = the computed 1-a lower confidence limit
b_2 = the computed 1-a upper confidence limit.

The upper and lower confidence limits a_1, a_2, b_1 and b_2 and functions of the desired confidence level and the sampling distribution of the sample statistic used to estimate ?.

one-sided confidence interval would be appropriate for comparing the observed abundance of an organism to the minimum abundance deemed necessary for survival of the species and/or for the health of the ecosystem. Because upper and lower 1-sided confidence intervals are easily derived from the corresponding 2-sided confidence intervals, only the latter will be presented in this appendix. In nearly all cases, the desired 1-sided confidence intervals can be obtained by substituting the appropriate upper and lower bound definitions from Box 2; primarily, this involves using 1-α levels of t-, z- or χ^2 statistics in place of 1-α/2 levels of the given two-sided formulae.

The general formulae in Box 2 can be used to compute two-sided 100×(1-α)% confidence intervals around sample estimates that are normally distributed (e.g., means, binomial proportions). Such confidence intervals are symmetric; i.e., the distance between the estimates and their upper or lower bounds, are equal. The distance between the estimate and its upper or lower bound is called the confidence interval half-width (W) and is a measure of the precision of the estimate; the smaller the distance the greater the precision. The size of the half-width depends on the size of the variance (S^2 for means and p(1-p) for proportions), the sample size and the specified confidence level. When each of the other two factors is held constant, the following changes will result in wider confidence intervals and thus less precise estimates of water quality:

1. increasing the variance
2. increasing the confidence level (e.g., going from 80% to 95% confidence)
3. decreasing the sample size

Specific formulae, and details for the construction of confidence intervals for a variety of population parameters are presented (with examples) in Appendix D.

C.1.6 Simple Random Sampling Designs

Having specified the tolerable error rates and the minimum sample size needed to assure them, the investigators can be reasonably confident that their resulting sample estimate(s) will meet their precision requirements. However, as Fig. 1a demonstrates, precision is not the only concern; the investigators must also insure that the sample estimates are unbiased (Fig. 1d). Assuming that sampling gear and laboratory procedures are not defective (i.e., negligible measurement error), bias can be controlled through proper implementation of an appropriate **sampling design**. Detailed treatment of this topic is available from the EPA document, *Guidance for Choosing a Sampling Design for Environmental Data Collection* (EPA QA/G5-S). In this section we present a brief overview of the sampling design process with an emphasis on the issue of representative sampling.

The first step in designing a sampling program is to construct the **sampling frame**. The sampling frame is simply a listing of all the sampling units in the target population. For example, consider a rectangular section of the benthos of a pond from which an estimate of the sediment concentration of pesticide X is desired. Assume further that the area has been divided into 160 equal-sized grid cells. Although the true population is an infinite number of points, the grid provides a convenient frame that completely covers the target area. Fig. 2a displays such a grid with the center of each cell indicated with either an open or a solid circle. The grid cells are the population elements and hence the potential sampling units. A sampling frame is constructed by uniquely identifying each of the 160 sampling units; e.g., by numbering them from 1-160, starting in the upper left corner. Next, we select (without replacement) numbers in the range of 1-160 from a Table of random numbers to identify which population members to include in the sample. A set of n=30 randomly selected grid cells and a second randomly selected set of n=10 are denoted by solid circles in Figs 1a and 1b, respectively. The mathematical theory of combinations and permutations states that when n sampling units are randomly selected from a population of size N, it will be possible to draw S different, but equally likely samples:

$$S = \frac{N!}{n!\left(N-n\right)!} \tag{1}$$

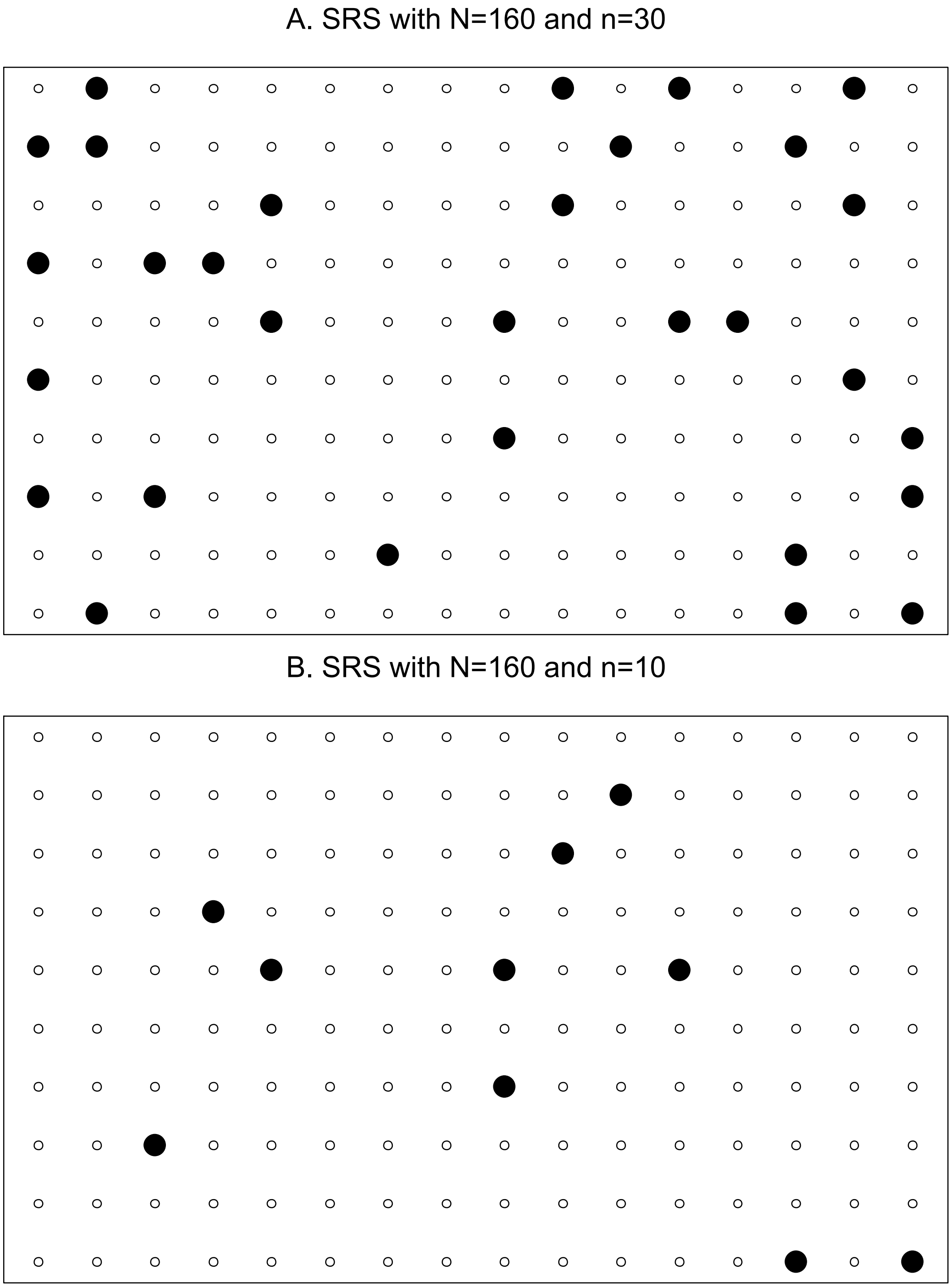
Thus one could draw 2.749×10^{32} samples, each of size 30, or 2.274×10^{15} samples, each of size 10, from the 160 member target population. A simple random sample (SRS) is one in which each of the S possible samples has an equal probability (i.e., 1/S) of selection. For the benthic target population in Fig. 2, this insures that selection will not be biased for or against any part of the benthic area; i.e., the simple random samples are unbiased. However, closer inspection of Figs 2a and 2b reveals that, although unbiased, the two samples don't provide the same amount of information. Whereas sample 2a (n=30) provides information on pesticide concentrations from all four corners and the center of the target area, Sample 2b (n=10) lacks information from three of the four corners of the target area. Thus an SRS of n=30 appears to provide reasonably good coverage of the target population, while an SRS of n=10, does not.

C.1.7 Representativness and Independence

The difference between the coverages of the two samples in Fig. 2 illustrates the concept of representativeness in the sampling of a target area. As illustrated, a **representative sample** is a sample, which, in microcosm, captures the range of the variability of the attribute of interest (e.g., the sediment concentration of pesticide X), among the elements of the target population. Note that unbiasedness in the sampling process does not necessarily insure representativeness of the resulting sample(s). While "representativeness" is relatively easy to conceptualize in a spatially referenced sampling frame such as Fig. 2, it is much more difficult to formulate an unambiguous mathematical definition.

A sampling design that draws only from a specific subset of the target area, will yield unrepresentative samples. Unrepresentative sampling often occurs as a result of selectively sampling from the more accessible locations in the target population. For example, it may be much easier to obtain benthic samples from under bridges than from other stream reaches in a heavily forested area. To the extent that the measured characteristic (e.g., abundance of chironomid larvae) differs close to roadways vs. within the forest, the result will be an estimate like that shown in Fig. 1a or 1c. For this reason such **convenience samples** tend *not* to be representative. Similarly, samples taken close together in space and/or time tend to be more alike (i.e., **autocorrelated**) than more widely dispersed samples. For example, a sample of 30 sampling units from a single cove in a lake will not contain as much information as a sample of 30 sampling units, each coming from a different cove. This is because samples of autocorrelated

Fig. 2. Two examples of simple random sampling (SRS) from a 10x16 grid. Each circle represents a spatially fixed sampling unit; solid circles are sampling units that were randomly selected for inclusion to samples of sizes 30 (A) or 10 (B).



A. SRS with N=160 and n=30

B. SRS with N=160 and n=10

sampling units contain redundant information. To the extent that 30 autocorrelated sampling units are alike, they may carry no more information than 1 or 2 **independent** (i.e., uncorrelated) sampling units. Thus, for the purpose of making inferences about the variability of the measured attribute over the entire lake, the **effective sample size** of the sample from the single cove may be as small as one or two sampling units, compared to 30 from the multiple-cove sample.

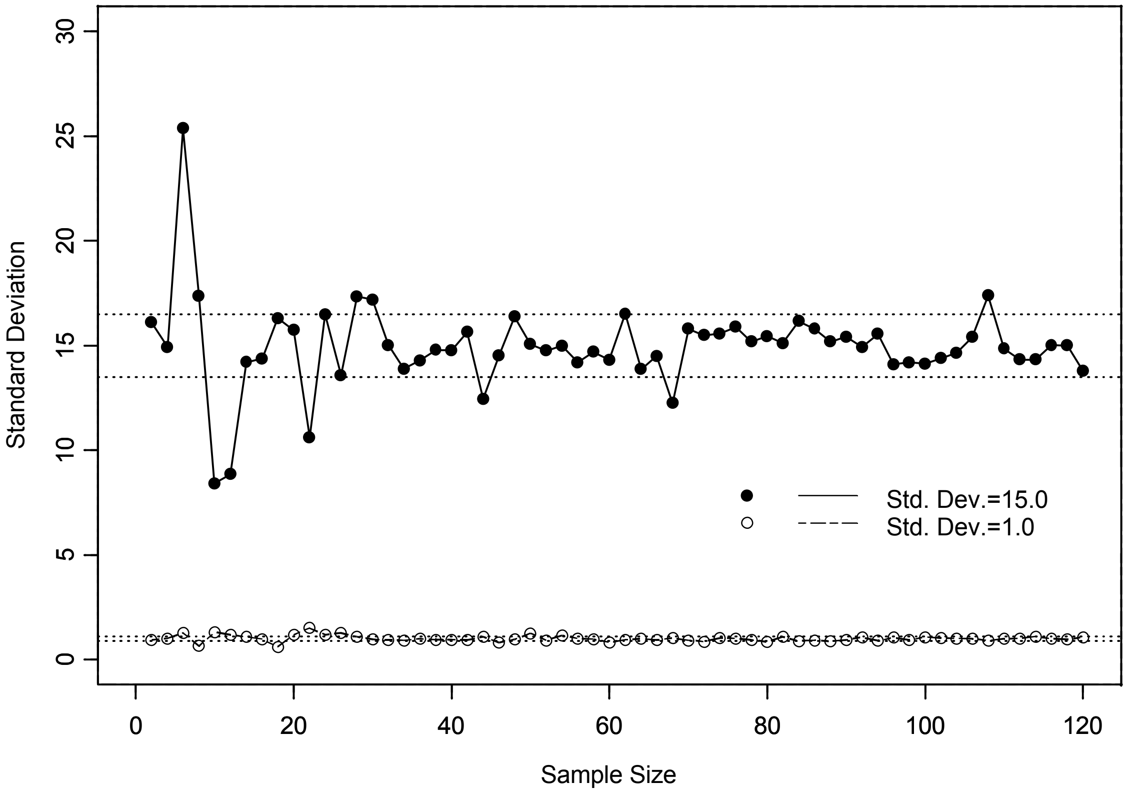
The problem of independence and sample size can be considered from variance component perspective. For example, a toxicologist may want to estimate the mean concentration of mercury in muscle tissue of a population of bass in a lake. Because of heterogeneity in the distribution of mercury exposure within the lake, heterogeneity of feeding behaviors, and physiological factors among individual bass, the tissue concentration of mercury can be expected to vary among fish within the population. If the sample mean is to be valid estimator of the mean of the population of bass in the lake, it must be computed from a representative sample of the bass target population. If the sample is truly representative, the variability (i.e., the heterogeneity) of the tissue concentration of mercury among bass in the sample will be an unbiased estimator of the mercury variability in the target population. Moreover, if the variability in the target population is high, the sample will need to be large (e.g., n=100 bass) in order to capture the variability of the population.

The among-bass variance in mercury concentration estimated from a probability sample of 100 fish selected from the lake population will be larger than a variance computed from 100 tissue samples taken from a single bass or from 100 bass taken from the same cove. In other words, neither the within-fish variability nor the within-cove variability will be the same as the among-fish variability for 100 different bass that were randomly selected from throughout the entire lake. In general, the heterogeneity of suprapopulations will be greater than that of component subpopulations because of the redundancies associated with correlations described above. Thus one should never limit sampling designs to subpopulations or subsets of the desired target population or target area.

The best way to ensure representativeness is to collect, using an unbiased selection method (e.g., SRS), a sufficient number of independent sampling units to capture the variability inherent in the target population. But how does one determine when the number of sampling units is "sufficient"? The best answer is, "conduct a pilot study". Using simulation techniques, Browne (1995) demonstrated that substituting the upper one-sided 1-α% confidence limit of the pilot sample standard deviation for S in Box 1a of Appendix D provided a sample size estimate sufficient to achieve the planned power (1-β), for a prespecified effect size in at least 1-α% of his Monte Carlo experiments.

Elliot (1977) described a more labor intensive (and more expensive) method in which a series of pilot samples with sample sizes increasing in increments of 5 sampling units is used to obtain a plot of the sample mean or variance against the sample size. The sample size at which the amplitude of fluctuation of the sample estimates damps indicates the minimum number of sampling units for a representative sample. This approach is illustrated in Fig. 3 in which sample sizes of single samples, increasing from n=2 to n=120, in increments of 2 sampling units, are plotted against the corresponding sample standard deviations. The open circles represent samples taken from a population with N=25,000, μ=20 and σ=1 while the solid circles represent

Fig. 3. Sample estimates of the population standard deviation plotted against the sample sizes of samples from two different populations. Open circles are estimates from a population whose true standard deviation is 1.0; solid circles are estimates from a population whose true standard deviation is 15.0. Broken lines enclose sample estimates that are within ± 10% of the true population value.



samples from a population with N=25,000, μ=20 and σ=15. Fluctuations in the standard deviations from the later population stabilize at about n=32, with sample estimates remaining within ± 10% of the population value (i.e., within the two broken lines), thereafter. Thus it appears that a sample of size n=32 is sufficient to represent the variance in the target population. Furthermore, increasing n from 32-120 does not appear to appreciably increase the representativness. By contrast, samples sizes as small as n=14 appear to quite representative of the population with σ=1. In general the less variable a population (i.e., the smaller σ), the smaller the required sample size. For example in an extreme case where every fish in a population of 100 is the same age, a sample of n=1 will be representative of the population age distribution.

C.1.8 Choosing a Sampling Design

The principle advantages of the SRS design are that (1) it provides statistically unbiased parameter estimates, (2) it is simple to understand and implement and (3) sample size calculation and subsequent statistical analyses of the sample data are straightforward. Simple random sampling is useful when the target population is relatively homogeneous; i.e., when there are no pronounced spatial or temporal patterns in the distribution of the population members or localized areas of extreme population abundance ("hot spots"). In those cases where SRS designs are not appropriate other probability-based sampling designs are available. Five of these, stratified sampling, systematic grid sampling, ranked set sampling, adaptive cluster sampling and composite sampling are reviewed in *Guidance for Choosing a Sampling Design for Environmental Data Collection* (EPA QA/G5-S). Each design has specific advantages in specific situations. Although they differ in details of selection of the sampling units, because they are all probability-based designs, they share two essential features: (1) each member of the target population has a known (though perhaps unequal) probability of selection into the sample; and (2) techniques of statistical inference (i.e., confidence intervals and hypothesis tests) can be applied to the resulting data. Data obtained from convenience or judgment sampling cannot be used to make formal statistical inferences unless one is willing to assume that they have the same desirable properties as probability samples, an assumption that usually cannot be justified (Peterson et al. 1999).

C.1.9 Data Quality Objectives Case History: Monitoring Dissolved Oxygen (DO) Downstream from an Agricultural Operation

This example concerns the use of a binomial proportion in planning for environmental decision-making. The example is presented in a continuous format to show the seven-step DQO Process in its entirety.

0. Background

Both flowing and standing waters contain dissolved oxygen (DO). Oxygen enters water principally from two sources: diffusion across the air-surface water interface and from photosynthesis of aquatic vascular plants and green algae. The capacity of water to hold oxygen depends on the complex interaction of several factors. Oxygen in water is negatively associated with temperature, altitude, salinity, plant and animal respiration, and decomposition of organic matter by microorganisms. Generally dissolved oxygen decreases steadily from dusk to dawn and increases during the daylight hours.

Because oxygen is required for respiration, aquatic community structure, composition and health are to a great extent determined by the dissolved oxygen content of the waters they reside in. Warm-water fishes require a minimum dissolved oxygen concentration of 5.0 mg/l, while cold-water fishes require a minimum of 6.0 mg/l. Sustained exposure to dissolved oxygen concentrations < 2.0 mg/l for four days or more will kill most of the biota in an aquatic ecosystem. Such systems quickly become dominated by chemosynthetic bacteria (e.g., foul smelling sulfur bacteria) and undesirable air-breathing invertebrates.

Inputs of large amounts of organic material such as fertilizers or raw sewage encourage rapid proliferation of decomposing bacteria. Decomposition of large organic molecules (e.g., starches, fats, proteins, etc.) to smaller molecules (e.g., methane) is primarily an oxidative process that can rapidly deplete DO in the water column. Depletion may be transient, subsiding when the organic material has been completely decomposed or washed away or it may result in the permanent replacement of one aquatic community with another if the organic input continues for long periods of time. Thus dissolved oxygen is an important parameter that should be closely monitored, especially in ecosystems that are exposed to the risk of allocthanous inputs of organic materials from agricultural or industrial operations or from human residential development.

1. State the Problem

How were the planning team members selected? The planning team included a state limnologist (the project manager), a regional fisheries biologist, an environmental scientist from the EPA regional office, and a consulting statistician. The decision makers were the limnologist and the fisheries biologist.

How was the problem described and a conceptual model of the potential hazard developed? The problem was described as monitoring the dissolved oxygen concentration at a downstream location from a commercial hog operation. The operation has several large waste lagoons that have experienced overflow and runoff problems in the last 5 years. There have been several fish kills in the last 3 years just downstream of a large swale through which runoff from the hog facility has escaped in the past. The operators have recently made improvements to the lagoon system and the landscaping that are supposed to remedy the problem. The state pollution control agency has made a decision to closely monitor dissolved oxygen downstream of the swale for the next 3 years to determine if the problem has been resolved.

The conceptual model consists of a single discharge point into the river (i.e., the location of the swale), a half-mile mixing zone immediately downstream of the hog facility, and a cross-section of the river at a point approximately 0.60 of a mile from the swale and just beyond the mixing zone. The effluent consists of runoff of high organic material from the manure lagoons. Based on widely accepted models of stream dynamics it is assumed that mixing of the effluent plume with the ambient waters is homogenous at distances greater than 0.50 miles downstream of the discharge. At these distances microbial activity supported by the organic effluent may deplete dissolved oxygen to dangerous levels throughout the water column. Direct measures of dissolved oxygen concentration (mg/l) taken along a cross-stream transect, located a half-mile downstream from the hog operation, were used to assess attainment of minimal DO standards.

What were the available resources and relevant deadlines? The state pollution control authority was prepared to commit a maximum of $5000 per year, for a 3-year period, to monitor dissolved oxygen concentrations in the vicinity of the hog operation. Monitoring was scheduled to begin in January 2000.

2. Identify the Decision

What was the Decision Statement? Consistent with existing state water quality standards, the decision statement was to determine if more than 10% of the DO estimates from regularly collected downstream samples fell below the 5.0 mg/l standard in any full calendar year of monitoring.

What were the alternative actions? If statistically significant nonattainment of the 10% DO standard was observed in any year during the 3-year monitoring period, the water would be listed and appropriate actions would be taken against the operators of the hog facility.

3. Identify the Inputs to the Decision

Identify the kind of information. Assessment of DO was made by direct *in situ* measurement of DO in a cross-section of the river approximately 0.60 miles downstream of the hog facility.

Identify the source of information. The state pollution control authority states that no more than 10% of the DO samples taken from a reach of river in a single year may be less than 5.0 mg/l.

What sampling and analytical methods were appropriate? DO measurements were made with a YSI Series 5700 DO probe, following procedures described in section 6 of the USGS National Field Manual for collection of water-quality data. Preparation, maintenance and calibration for the YSI 5700 were carried out following specifications at:

http://water.wr.usgs.gov/pnsp/pest.rep/sw-t.html

Dissolved Oxygen readings were recorded in the field. The results provided information on DO concentration at the monitoring site on each sampling date. The probe detection limit was well below the 5.0 mg/l action level.

4. Define the Boundaries of the Study

What population was sampled? Water passing through a cross-section of the river located 0.60 miles downstream from the hog facility.

What were the spatial boundaries? The spatial boundaries of the monitoring site were defined by a transect dawn perpendicular to the river, between two permanent markers each located 0.60 miles from the commercial hog operation. The transect spanned the width of the river and was 75 meters across. The depth along the transect between 5 and 70 meters from the left bank ranged from 2.6 to 3.7 meters with a mean of 3.1 meters.

What was an appropriate time frame for sampling? Sampling was scheduled to commence in January of 2000 and run through December of 2002.

What were the practical constraints for collecting data? There were no specific practical constraints in collecting the data within the specified time frame at the downstream sampling station.

What was the scale of the decision-making? The area-adjusted mean DO computed from DO measurements taken along the transect on a specific sampling date was the basic unit of analysis. The area-adjusted mean represents the DO in the cross-section of the stream lying beneath the transect on a specific sampling date.

5. Develop a Decision Rule

What was the decision rule and Action Level? Because of its importance to aquatic community structure and health, state DO criteria permit no more that 10% of the samples taken from a body of water during a year the have less than the 5.0 mg/l minimum. This proportion was computed by comparing each area-averaged mean DO to the standard and dividing the number of nonattainment values by the total number of sampling times in a year. The 10% criterion is the action level. If the true proportion of nonattainments exceeded the action level, remedial action and listing of the river was indicated.

6. Specify Tolerable Limits on Decision Errors

How was the baseline condition set? The baseline condition assumed by the investigators was that the water in the cross-section attained the DO criterion is each year in which it was monitored. This decision was made because the baseline is traditionally assumed to be attainment. However in this case, the choice of the baseline was trivial because the investigators intended to specify balanced false-acceptance and false-rejection error rates.

How was the gray region specified? The gray region was designated by considering the consequences of distinguishing between 10% and 25% nonattainment. The benefits of distinguishing among relatively small exceedance rates within the gray region were deemed insufficient to justify the cost (Table 1). Thus, following Smith et al. (2000) the upper bound of the gray region was set at 25% nonattainment.

How were tolerable decision error limits set? The consequences of not detecting low DO events need to be balanced with the consequences of falsely declaring that the reach of the river immediately downstream of the hog operation had seriously depleted dissolved oxygen. The consequences of the former include fish-kills and undesirable community restructuring. The consequences of the latter are unnecessary economic hardship on the operators of the commercial hog facility and perhaps incorrectly listing the river. In this case, it was decided that the two types of error were equally undesirable. Thus, a decision was made to simultaneously minimize the false negative and false positive error rates to the same value. Given the fiscal constraints on the monitoring program, it was determined that the common error rate associated with a gray region of width 0.15 should be no greater than 0.15. This is illustrated in Figure 4 and Table 2.

7. Optimize the Design for Obtaining Data

What was the selected sampling design? Following the recommendations in Section 6.0.2B of the USGS National Field Manual for collection of water quality data, DO measurements were made *in situ* along a fixed transect using the Equal-Width Increment method (EWI). The method calls for the transect to be divided into at least 15 increments of equal width. DO measurements were made in the middle of each interval of the transect at the middepth of the vertical between the surface and the stream bottom. The surface-bottom depth was recorded along with the DO in each interval. The cross-sectional area of each increment was computed as the product of the surface-bottom depth and width of the interval. The total cross-sectional area of the stream at the sampling station was computed as the sum of the areas of the all of the intervals. Finally, the area-adjusted mean DO of the cross-sections of the river at the sampling station was computed as:

$$\overline{DO} = \sum_{i=1}^{n} w_i DO_i \qquad (1.1)$$

where w_i = the cross-sectional area of the i^{th} interval divided by the total cross-sectional area of the stream at the sampling station

DO_i = the middepth DO concentration at the center of the i^{th} interval of the transect

n = the total number of intervals that the transect was partitioned into.

The distance of the transect from shore-to-shore at the sampling station was 75 meters. The transect was divided into 15 increments, each of 15 meters in width. Based on the cost constraints, the width of the gray region, and the 15% balanced error rates, it was determined that DO should be measured at the sampling station at 11-day intervals beginning on January 3 of each year and ending on December 31, for a total of 34 evaluations per year (@ $150 per evaluation), for an annual cost of $5100. It was decided that the budget should be adjusted to permit the small ($100) cost overrun.

Fig. 4. Decision performance goal diagram for a *Z*-test based on the normal approximation to the binomial distribution. H_0: population proportion of DO concentrations that is ≤ 5.0 mg/l, is ≤ 0.10 vs. H_a: population proportion > 0.10, when the sample size=33 and the width of the gray region is 0.15 (minimum detectable effect size=d=0.15).

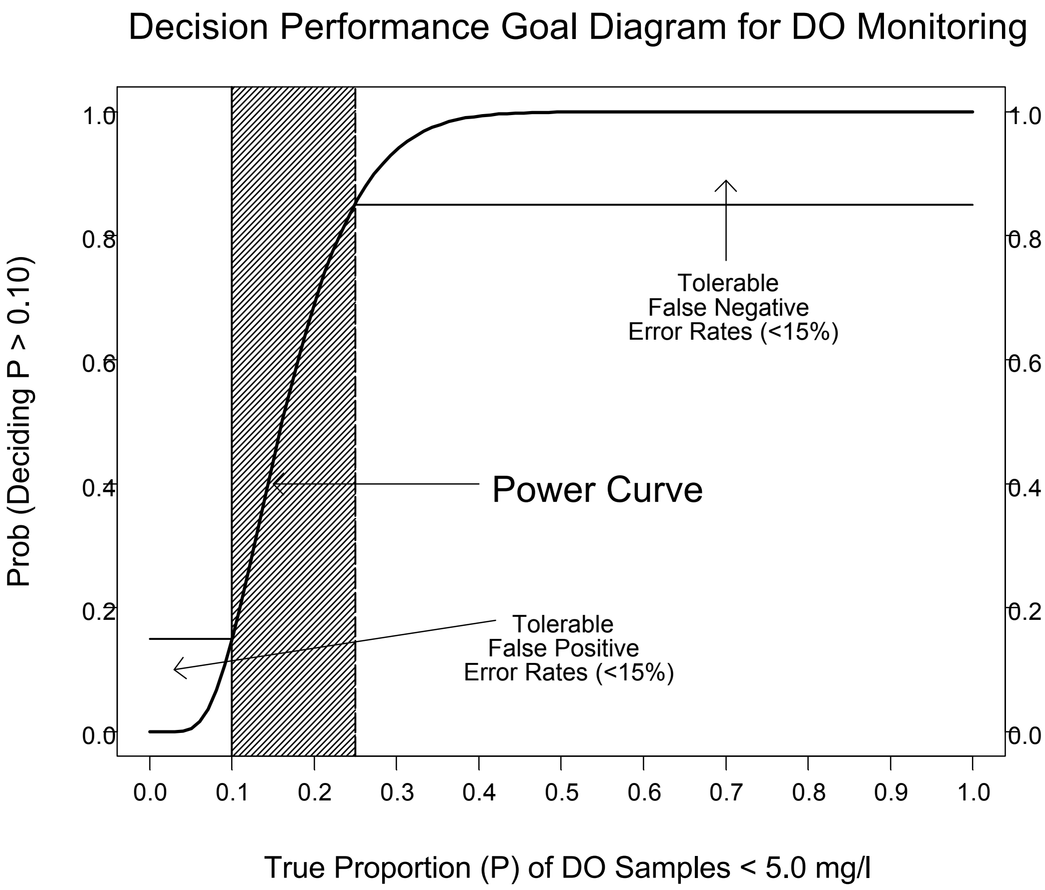


TABLE 1. EFFECTS OF INCREASING THE WIDTH OF THE GRAY REGION ON SAMPLE SIZE AND COST

WIDTH OF GRAY REGION	FALSE POSITIVE RATE	FALSE NEGATIVE RATE	SAMPLE SIZE	COST
0.01	0.15	0.1500	4036	$605,400
0.05	0.15	0.1494	186	$27,900
0.10	0.15	0.1485	53	$7,950
0.15	0.15	0.1473	26	$3,900

TABLE 2. EFFECTS ON THE SAMPLE SIZE AND COST OF INCREASING BALANCED ERROR RATES

FALSE POSITIVE RATE	FALSE NEGATIVE RATE	SAMPLE SIZE	COST
0.05	0.0481	65	$9,750
0.10	0.0983	40	$6,000
0.15	0.1473	26	$3,900

What were the key assumptions supporting the selected design? Four assumptions were required to support the sampling design:

1. Flow velocity was relatively uniform along the transect
2. Stream depth was relatively uniform along the transect
3. There was homogenous mixing of any effluent from the hog operation with the ambient waters between the discharge point and the sampling station
4. The 11-day sampling interval was sufficiently small that no significant low-DO events would be missed

Assumptions 1 and 2 were verified by direct measurement along the transect prior to initiation of the study. Assumption 3 was accepted based on numerous published limnological studies of mixing dynamics in rivers. The 4th assumption represents a compromise between the desire for good coverage of the time period of concern and the project cost constraints. Although short-term transient episodes of oxygen depletion might be missed with an 11-day sampling interval, it was decided that the cost of more frequent sampling could not be justified given that short-term depletions severe enough to cause significant fish-kills would be observed and reported in any event. A systematic sample of the kind used in this study was deemed sufficient to capture patterns of "typical" fluctuations in DO during a 3-year period at the monitoring site.

C.2 The Data Quality Assessment Process: Exploratory Data Analysis

C.2.0 Review of the Steps in a Basic DQA

The data quality assessment (DQA) process is described in detail in the EPA document, *Guidance for Data Quality Assessment EPA QA/G-9* (EPA/600/R-96/084). EPA's DQA process is a 5-step plan designed to provide scientifically defensible conclusions through statistical analyses of environmental data previously collected in accordance with a well-conceived DQO strategy. The DQA steps are:

1. Review the study DQOs and associated sampling design(s)
2. Conduct a preliminary review of the data
3. Based on the data and the research question, select an appropriate statistical test(s)
4. Verify the underlying assumptions of the selected test(s)
5. Draw conclusions from the data analyses

Activities in DQA Step 1 should focus on:

(a) Review of the outputs of the four statistical DQOs (i.e., target population definition, decision rule(s), acceptable limits of decision error, and sampling design)
(b) Translation of the decisions identified in the DQOs into statistical hypotheses
(c) Confirmation of the limits on the decision error rates
(d) Any features of the sampling design that would bear on the selection or interpretation of the statistical test(s)

The remainder of this Appendix will focus on techniques and considerations that are useful for carrying out steps 2-5 of the DQA process. We begin with a discussion of a graphically based approach to Step 2 which coincidentally contributes to the resolution of many of the issues associated with the subsequent steps in the DQA process. In Sections C.3.0 – C.3.3 we will review the basic principles of statistical hypothesis testing with emphasis on those aspects that bear on control of decision error rates. Various types of statistical tests appropriate for water quality assessments will be reviewed and some examples of their application and interpretation will be presented in Appendix D.

C.2.1 Exploratory Data Analysis: Basic Principles

The second DQA step involves an initial quality assurance (QA) review of the data for data entry errors, etc. This is followed by computation of the statistics specified in the DQO process and/or any auxiliary statistics that the analyst might deem necessary to the interpretation of the data or to the assessment of the assumptions of the statistical tests. Finally, and perhaps most importantly, step 2 requires extensive **exploratory data analyses (EDA)**. EDA usually involves graphical methods that facilitate identification of characteristics and/or relationships in the data that are often crucial to proper interpretation of the subsequent statistical tests. Indeed, the EDA results may suggest that additional or different tests should be applied (Cleveland 1993), or that some transformation of the data may be necessary. Graphical EDA methods are described in detail in *Guidance for Data Quality Assessment EPA QA/G-9* (EPA/600/R-96/084), *Biological*

Criteria: Technical guidance for Survey Design and Statistical Evaluation of Biosurvey Data (EPA 822-B-97-002*)*, and in *Visualizing Data* (Cleveland 1993). These techniques are easily implemented with standard commercial statistical software (e.g., SASGRAPH, SPLUS, SPLUS EnvironmentalStats and SPLUS Spatial).

It was pointed out in Section C.1.5 that some confidence interval estimators require that the sampling units and/or the sampling distributions be approximately normally distributed. Similar assumptions of normality will be required for the validity of many of the statistical tests in Appendix D. In Section C.1.7, additional requirements for spatial and temporal independence among the sampling units were imposed for the estimation of confidence intervals. The independence requirements also apply to the statistical hypothesis tests that are discussed in Appendix D. There are several ways to assess the validity of these assumptions but the most effective methods involve graphical EDA procedures (Cleveland 1993). Examples of graphical verification of normality and of graphical assessment of spatial and of temporal independence will be illustrated in this section of the appendix.

C.2.2 EDA Example 1: Assessing Normality of Continuous Data

Two hundred and forty-four turbidity readings (Table 3) were recorded from 1980-2000 in a reach of the Mermentau River in Southwest Louisiana. From the sample, the investigators computed the following descriptive statistics:

n=244
Minimum = 8
Median = 75.5
Maximum =600
Mean=100.3
Std. Deviation =86.1

The investigators wanted to determine if the mean turbidity was greater than the local criterion value of 150 NTU. Two options were available to them: (1) they could compute an upper 1-sided 95% confidence interval (Appendix D, Box 3) on the mean turbidity and check to see if 150 was included in within the interval or (2) they could compute a 1-sided t-test (Appendix D, Box 8). Both methods require that the distribution of the 244 turbidity readings be approximately normal. The most important attributes of the normal distribution are that it is symmetric and that its mean and median are equal. A quick examination of the turbidity descriptive statistics reveals that these conditions do not hold for the sample, suggesting that the turbidity data are not normal.

The best way to assess the form of a population distribution is to graph its frequency distribution. A frequency distribution is a histogram that displays the way in which the frequencies (i.e., counts) of members of the sample are distributed among the values that they take on. The corresponding **relative frequency distribution** (Fig. 5A) can be calculated by dividing each count by the total sample size. The familiar “bell curve” is a graph of the expected relative frequency distribution of a variable that is normally distributed (e.g., heights or weights of men or women). When the relative frequency is based on smaller sample sizes, it will take the form of a histogram composed of clearly discernible individual bars; when it is computed from very large numbers of individuals (e.g., hundreds of thousands of women) its graph tends to smooth out as the bars increase in number and merge together to produce a smooth surface (i.e., the bell).

If the counts of the individuals are large enough, it is often possible to summarize the relative frequency distribution with a mathematical expression called the **probability density function** (PDF). The PDF predicts the relative frequency as a function of the values of the random variable and one or more constraining variables, called model parameters, which can be estimated from the sample data. Continuous distributions whose PDFs can be so defined are called **parametric continuous distributions**. The PDF is the algebraic expression for the line that delimits the shape of the relative frequency distribution. In Fig. 5B, the plot of a PDF for a normal distribution (bell-curve) is superimposed on the relative frequency distribution of the sample of turbidity values (Table 3) from which it was computed. The specific form of the bell curve is controlled by the 2 parameters of the normal distribution: the population mean (μ) and the population standard deviation (σ). The larger the mean, the father the center of the bell is shifted to the right; the larger the standard deviation, the wider and lower the bell.

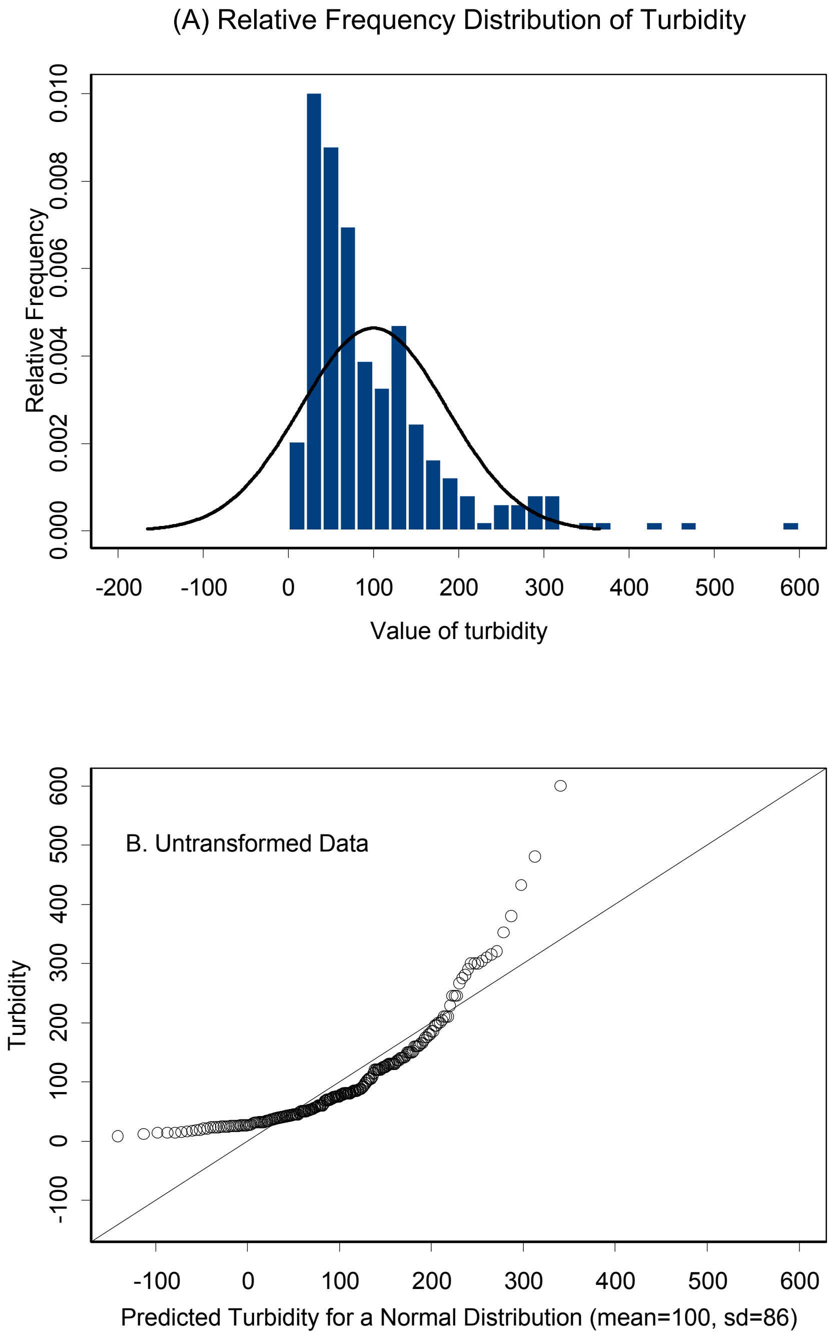
In Fig. 5A, the turbidity data have been divided into 21 groups, each of which has a vertical bar associated with it. The height of the bar indicates the proportion of the 244 turbidity measures that are in the group, which in turn, is an estimate of the probability of any turbidity measure taken from the river reach being in the group. The bell curve is based on the assumption that the actual data come from an underlying normally distributed population with μ= the sample mean=100.3 and σ =the sample standard deviation=86.1. It is clear from Fig. 5A that the observed relative frequency distribution does not fit the assumed normal distribution very well. Specifically, the normal distribution predicts a substantial number of negative turbidity values, far fewer turbidity values in the range of 20-80 NTU than were actually observed and no values > 360.

Although Fig. 5A employs techniques (i.e., histograms and bell-curves) that are familiar to most scientists, it has the drawbacks that it requires significant effort to produce and that subtle differences between the assumed normal and the sample distributions can be obscured by the user’s choice of the number of groups into which to divide the data (21, in this case). Figure 5B, called a Q-Q-plot or a normal probability plot, is a much simpler graph that can be quickly and accurately read and interpreted. The Q-Q-plot is a graph of the sample data values (vertical axis) against the values predicted by the normal probability curve (horizontal axis), with μ and σ equal to the mean and standard deviation of the sample data. If the sample data exactly matched the normal predicted values, all the data points would lie on the diagonal line. In Fig. 5B, data points corresponding to the upper and lower tails of the sample distribution clearly deviate from normality. Because the deviant points are above the diagonal, it is clear that the tail values are considerably larger than would be expected for a normal distribution. Furthermore, it is easy to see that the turbidity data in the center of the distribution (i.e., 80-260 NTU) have values that are close to normal, but are slightly smaller than expected (i.e., they are located just below the diagonal).

Table 3. Monthly turbidity values (1980-2000) for a reach of the Mermentau River in Southwest Louisiana. Data are sorted by NTU (Nephelometric Turbidity Unit) value. Note that 232/244 sample values are less than 300 NTU; thus there are only 12 values between 300 and the maximum value of 600 NTU, indicating a long-tailed, highly skewed distribution.

8	12	14	14	14
15	16	17	18	19
21	21	23	23	23
23	24	24	24	24
25	25	25	25	25
26	26	26	26	26
27	27	28	30	31
31	31	32	32	32
32	32	33	34	34
34	35	36	36	37
37	38	38	39	39
39	40	40	40	41
41	41	42	42	43
44	44	44	45	45
45	45	45	45	48
48	50	50	50	50
50	50	50	52	52
53	53	53	54	54
54	56	56	57	57
58	60	60	60	60
60	60	62	64	68
68	70	70	70	**70**
70	71	72	72	74
74	74	75	75	75
75	**75**	**76**	78	78
78	78	80	80	80
80	80	80	80	80
80	81	82	83	85
85	85	85	85	86
87	88	88	88	90
91	93	95	98	98
102	102	105	105	105
105	108	114	115	120
120	120	120	120	120
120	123	123	125	125
125	126	128	128	130
130	130	130	130	130
130	133	136	136	136
140	140	140	140	142
145	150	150	150	150
150	152	160	160	160
160	162	165	165	170
175	175	180	180	185
185	195	195	200	200
204	210	210	210	228
245	245	245	266	275
280	290	300	300	300
304	310	315	320	352
380	432	480	600	

Fig. 5. Distribution of 244 monthly turbidity measurements (NTU) from the Mermantau River, 1980-2000. (A) Relative frequency histogram and normal probability function (μ = 100.3, σ = 86.1). (B) Normal Q-Q plot comparing distribution of the turbidity data to their expected values (diagonal) under the assumption that they are normally distributed with μ = 100.3 and σ = 86.1.



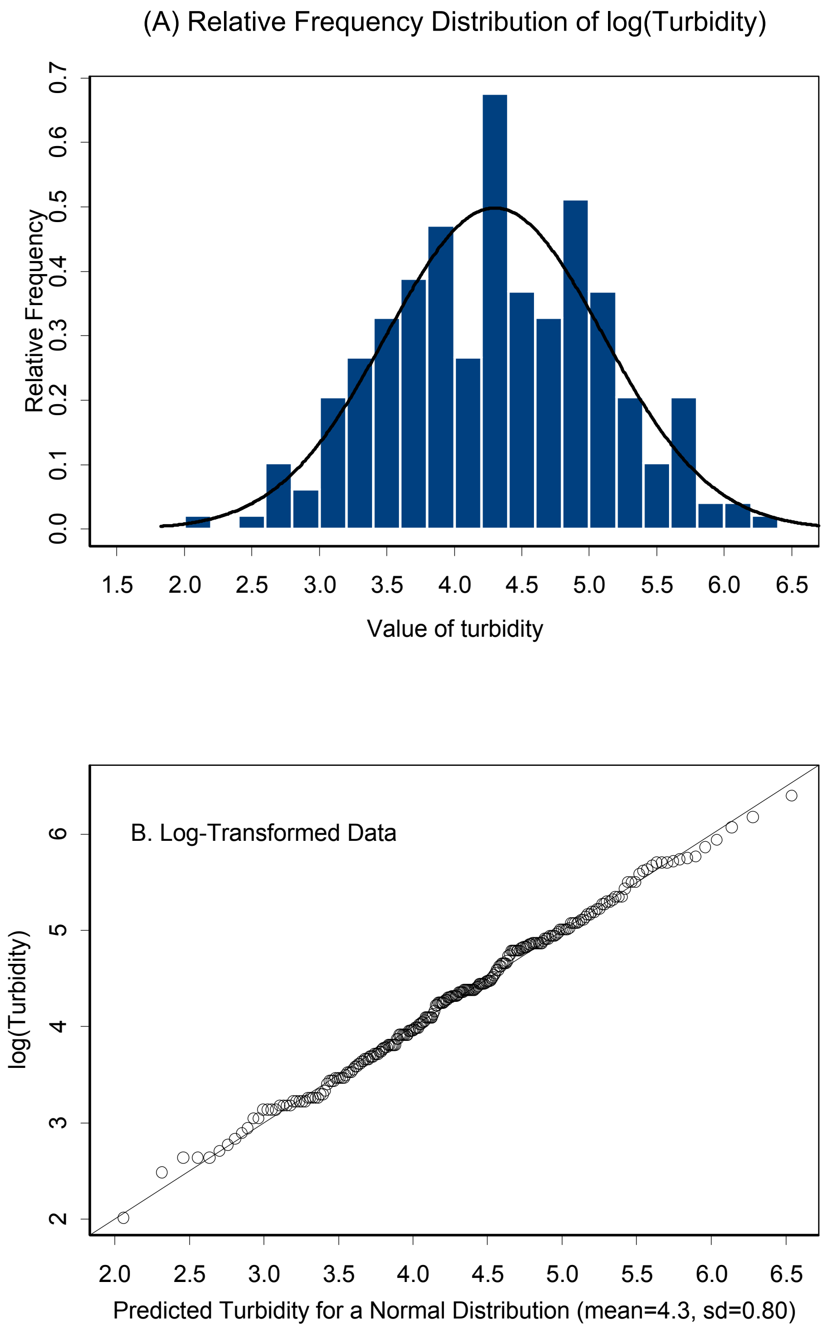
Normal probability plots are available from most commercial statistical software and can be produced with a single mouse click or a line or two of simple programming code. Alternatively, a number of formal statistical tests for the normality assumption are available from the same software packages. However, in practice, these tests (called goodness-of-fit tests) are not as useful as the graphical EDA methods. As with all such tests, they generate p-values that tell the user very little about the actual shape of the data distribution. Moreover, the GOF-test results are strongly dependent on the sample size. The normal GOF tests tend to accept the normality assumption for small sample sizes and reject it for large sample sizes, irrespective of the actual shape of the distribution of the sample data. The reasons for this troublesome behavior have to do with sample size and power relationships that are explained in Section C.3.2.

Having determined from the Q-Q plots and the sample descriptive statistics that the turbidity data are not normal, two courses of action are available: (1) find another test-statistic or confidence interval estimator which does not require normality or (2) apply a suitable transformation to the data such that the transformed values will be approximately normally distributed. Discussion of the first alternative will be taken up in section C.2.3. Here we will demonstrate how to apply the log-transformation and evaluate its effectiveness.

As noted, the distribution of the turbidity data is concentrated in the range of 0-200 with a very long tail of large values extending out to 600 NTU. This type of left-skewed, asymmetric distribution is very commonly observed for environmental and ecological variables. Influenced by the large values in the right tail, the sample mean of such a distribution will usually be substantially larger than the sample median (e.g., 100.3 vs. 75.5 for the turbidity data). A left-skewed distribution of positive-valued environmental data with mean > median is usually log-normally distributed (i.e., the distribution of logarithms of the raw data values is normally distributed). Fig. 6A displays the distribution of the logs of the turbidity data, overlaid with a plot of a bell-curve with μ= the log-scale sample mean=4.30 and σ =the log-scale sample standard deviation=0.80. The corresponding Q-Q-plot confirms that the turbidity data are approximately lognormally distributed.

There is an obvious problem in using log-transformed data for water quality attainment decisions; i.e., water quality standards are measured on the original scale. However, because of some fortuitous relationships among the means and medians of normal and lognormal data, this does not present a serious difficulty. Statistical inference based on the means of normal data depends on the fact that the mean and the median of normally distributed data are equal. Thus when a difference of say k-units is demonstrated between the mean of a normal distribution and a criterion value, one can assume that the center of the distribution also differs from the criterion by k-units. The back-transform of the log-scale mean [e.g., exp(4.3) for the turbidity data] of a log-normally distributed variable is called the geometric mean and is a close approximation to the median of the data on the original untransformed scale. Therefore, inferences based on the geometric mean of a lognormal distribution are equivalent to inferences based on the mean of a normally distributed variable. Comparison of the geometric mean to the criterion value should be interpreted no differently than a comparison involving the mean of a normal distribution.

Fig. 6. Distribution of 244 monthly log-transformed turbidity measurements from the Mermantau River, 1980-2000. (A) Relative frequency histogram and normal probability function ($\mu = 4.3$, $\sigma = 0.80$). (B) Normal Q-Q plot comparing distribution of the log-turbidity data to their expected values (diagonal) under the assumption that they are normally distributed with $\mu = 4.3$ and $\sigma = 0.80$.



The equivalence of the geometric mean and the median of the turbidity data can be easily demonstrated. The nonparametric estimate of the median turbidity and its 90% confidence limits (see Section D.1) is:

75.5(70.0, 80.0)

Noting that the Z associated with a 90% 2-sided confidence interval is 1.645 and that the log-scale mean and standard deviation are 4.3 and 0.80, we can apply the first equation in Box 1 of Appendix D (substituting z for t) to form a 2-sided 90% confidence interval on the log-scale mean:

4.301 (4.217, 4.385)

To obtain the geometric mean and its 90% confidence interval we simply exponentiate each of the values in the preceding expression:

73.8 (67.8, 80.2)

The geometric mean and its confidence interval differ only very slightly from the median estimate. The small disparity this is due to the fact that the distribution of the turbidity values is only approximately lognormal (Fig. 6).

C.2.3 EDA Example 2: Assessing Normality of Count Data

We have seen that the log-transform is appropriate for normalizing data for a skewed **continuous random variable**. However, it is also common to encounter distribution of skewed **discrete random variables** in environmental and ecological sampling; e.g., the distribution of snail counts among sampling units taken from the littoral zone or of exceedances in a collection of 1-liter sampling units from a lake. Very often the frequency distribution of such counts will approximate a Poisson distribution. Samples from a Poisson population are easily recognized by the fact that their mean and variance are equal or nearly so. Statistical theory insures that the distribution of the square roots of Poisson-distributed data is normal. Thus, the square-root transformation is recommended for normalizing skewed count data.

Roadside counts of meadowlarks from a survey of roads in the agricultural region of Southwestern MN in 1990 are presented in Table 4. The survey was designed so that an observer, driven over ninety different 5-mile stretches of road in the target area, counted the number of meadowlarks that were visible from the vehicle. In this design, each 5-mile stretch is a sampling unit, and the individual counts are the responses measured on them. The data were used as part of a long-term monitoring program to assess trends in the mean of the counts. The following descriptive statistics were computed from the 1990 sample data:

n=90
Minimum = 2
Median = 38
Maximum =143
Mean=46.1
Std. Deviation =36.2

Table 4 Meadowlark Counts

2	3	4	4	4
5	6	7	8	9
10	10	10	10	12
12	12	14	14	14
16	16	17	17	19
19	20	21	21	22
22	22	23	23	25
27	29	30	30	31
32	35	36	37	38
38	**38**	40	41	41
42	45	45	47	48
50	51	51	52	53
54	55	56	57	57
58	59	62	72	73
73	74	79	82	89
92	95	95	98	99
102	103	107	108	109
111	128	135	141	143

The relative frequency distribution of the counts, and an overlaid normal curve for a population with $\mu = 46.1$ and $\sigma=36.2$ are plotted in Fig. 7A. As suggested by the fact that the sample mean (46) is nearly 28% larger than the median, the distribution of counts in the sample is skewed. Similar to turbidity data, the Q-Q plot of the meadowlark counts (Fig. 7B) demonstrates that the most serious departures from normality occur in the tails of the distribution. However, the upper tail of the bird-count distribution, though longer than expected for a normal distribution, is much shorter than the upper tail of a lognormal distribution (compare to Fig. 5A). This is typical of Poisson data.

The results of the square-root transformation are displayed in Figs 8A and 8B. Although some lack-of-fit is still evident in the tails of distribution, the Q-Q plot indicates that the overall fit of the transformed counts to the normal distribution is quite good. As was the case with the log-transformed data, the back-transform of the mean of the square roots (i.e., 6.2^2) provides a close approximation to the median of the original data (i.e., 38.0). However well the square-root transform seems to have worked in this case, it is noteworthy that the mean and variance of the untransformed data are far from equal. This would seem to contradict the earlier statement that equality of the mean and the variance was an indicator of “Poissonness” and a justification for employing the square-root transformation to normalize the data.

Actually, this example illustrates the robustness of the square-root transformation. The major benefit of the log and square-root transformation is that they lead to an estimate of the median of the original distribution. But, it turns out that mean and median of almost any symmetric distribution will be equal or nearly so. Thus a transformation applied to a skewed distribution does not have to normalize it, it only needs to make it roughly symmetric. This is a less demanding requirement than normalization, hence the square-root and log transformations tend to work quite well on many skewed distributions even if they are not quite Poisson or lognormal. So, in practice, it is worth applying one or the other of the two transforms to all skewed data as a matter of routine, followed by verification of the results with Q-Q plots. If one of these transformations does not correct the skew, one of the nonparametric methods described in Section D.3 should be considered.

C.2.4 EDA Example 3: Assessing Spatial Independence

The concept of statistical independence among sampling units and its importance were introduced in Section C.1.7 of this appendix. Spatial *and* temporal independence are required for all of the inferential procedures described in this appendix. The following example illustrates the nature of the problem of spatial autocorrelation and its diagnosis; Example 4 will illustrate graphical analysis of temporal autocorrelation. Although counts of organisms are analyzed in the following example, the graphical diagnostics for the effects of spatial autocorrelation are equally applicable to spatially correlated distributions of chemical (e.g., pesticide residue concentrations) and/or physical variables (e.g., Secchi depths) in a target water. Moreover, the issues discussed here apply equally to sampling designs based on grids, transects, riffles or any other sampling units or clusters.

Fig. 7. Distribution of 90 meadowlark counts from SW Minnesota, 1990. (A) Relative frequency histogram and normal probability function ($\mu = 46.1$, $\sigma = 36.2$). (B) Normal Q-Q plot comparing distribution of the count data to their expected values (diagonal) under the assumption that they are normally distributed with $\mu = 46.1$ and $\sigma = 36.2$.

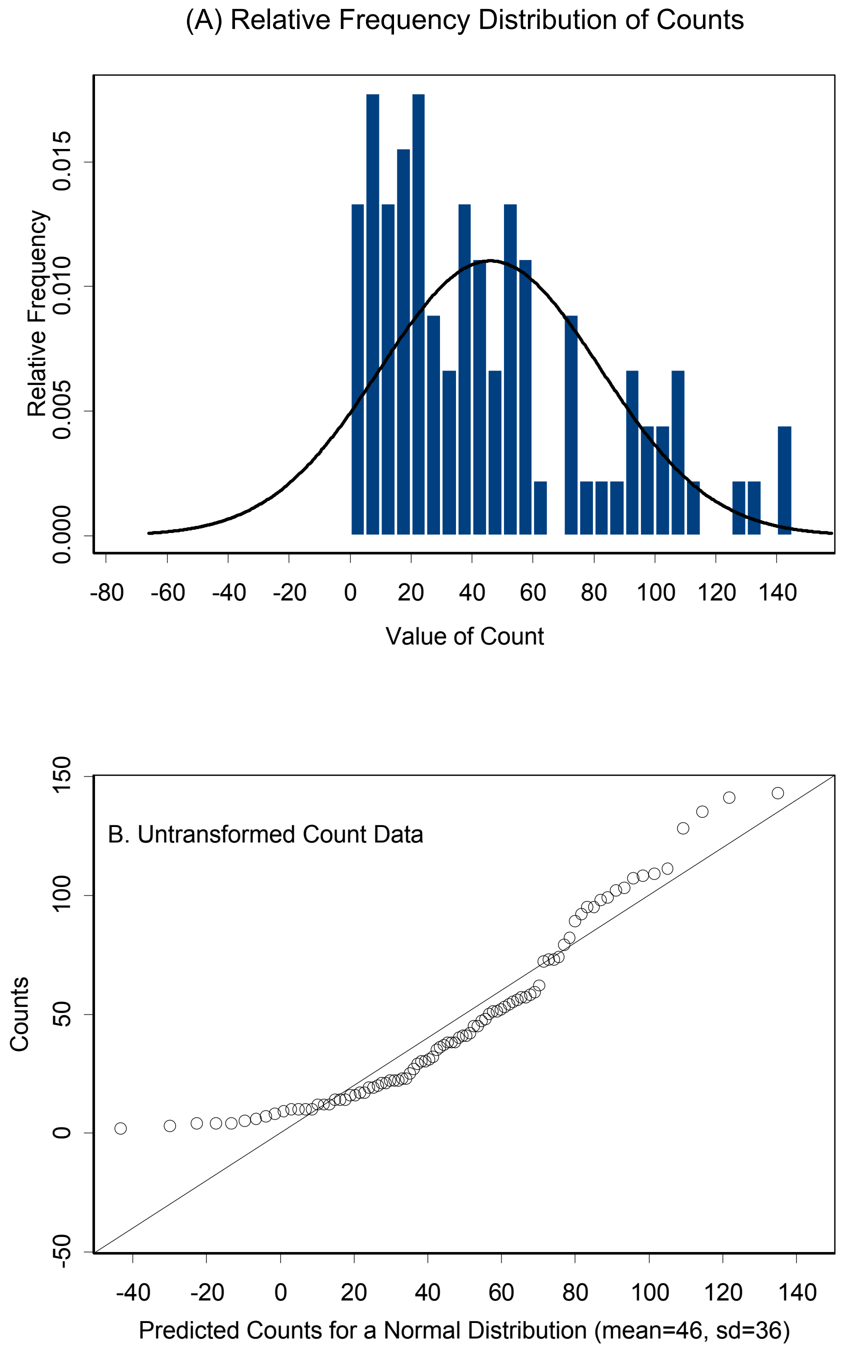
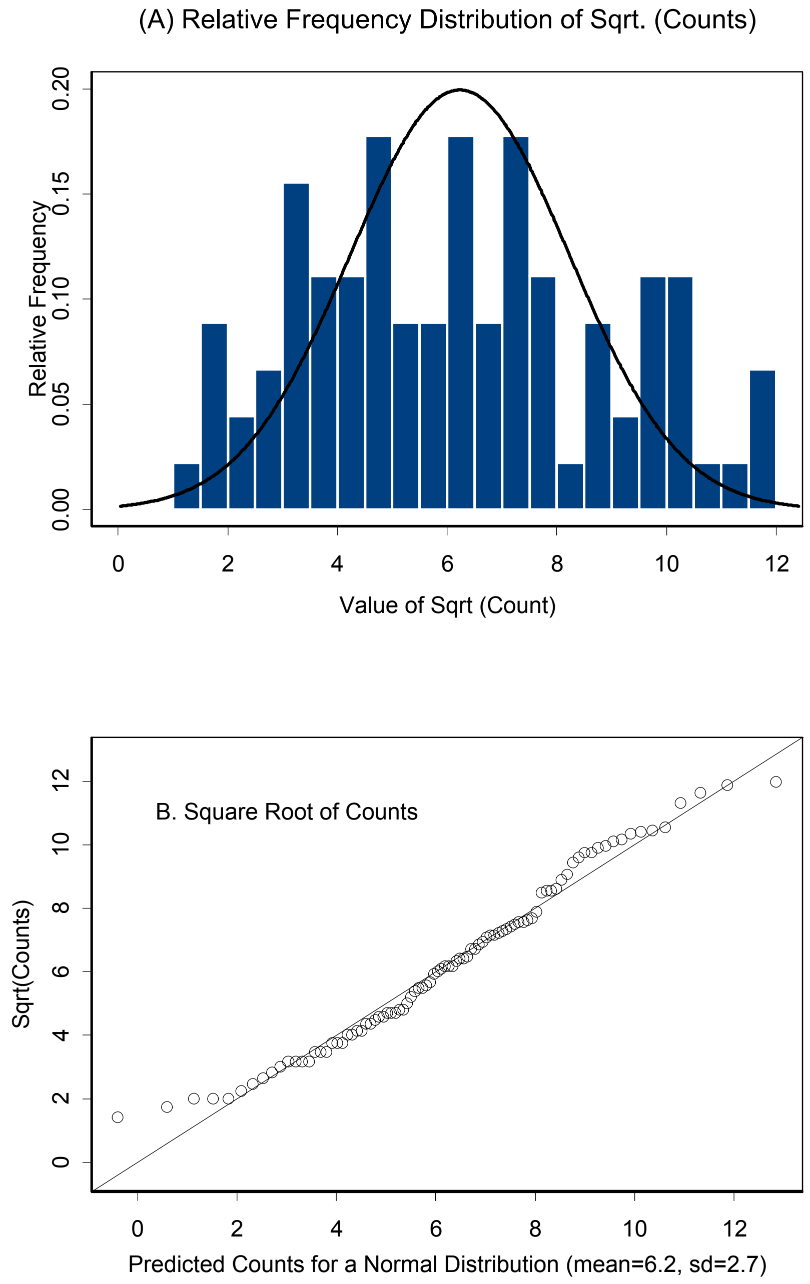


Fig. 8. Distribution of square roots of 90 meadowlark counts from SW Minnesota, 1990. (A) Relative frequency histogram and normal probability function ($\mu = 6.2$, $\sigma = 2.7$). (B) Normal Q-Q plot comparing distribution of the square roots to their expected values (diagonal) under the assumption that they are normally distributed with $\mu = 6.2$ and $\sigma = 2.7$.



Hengeveld (1979) conducted a study on the ecology of two species of beetle in a coastal flood plain. The study site was grided into a 21×12 rectangle of square cells, each of which was 40m on a side. Counts of each species were made once in the spring and once in the fall of 1975. In this design, each of the 252 grid cells is a sampling unit and the individual beetle counts are the measured responses. For this example we will only consider the spring counts of one of the species, *Dychirius globosus*. Basic descriptive statistics for the spring sample are presented below.

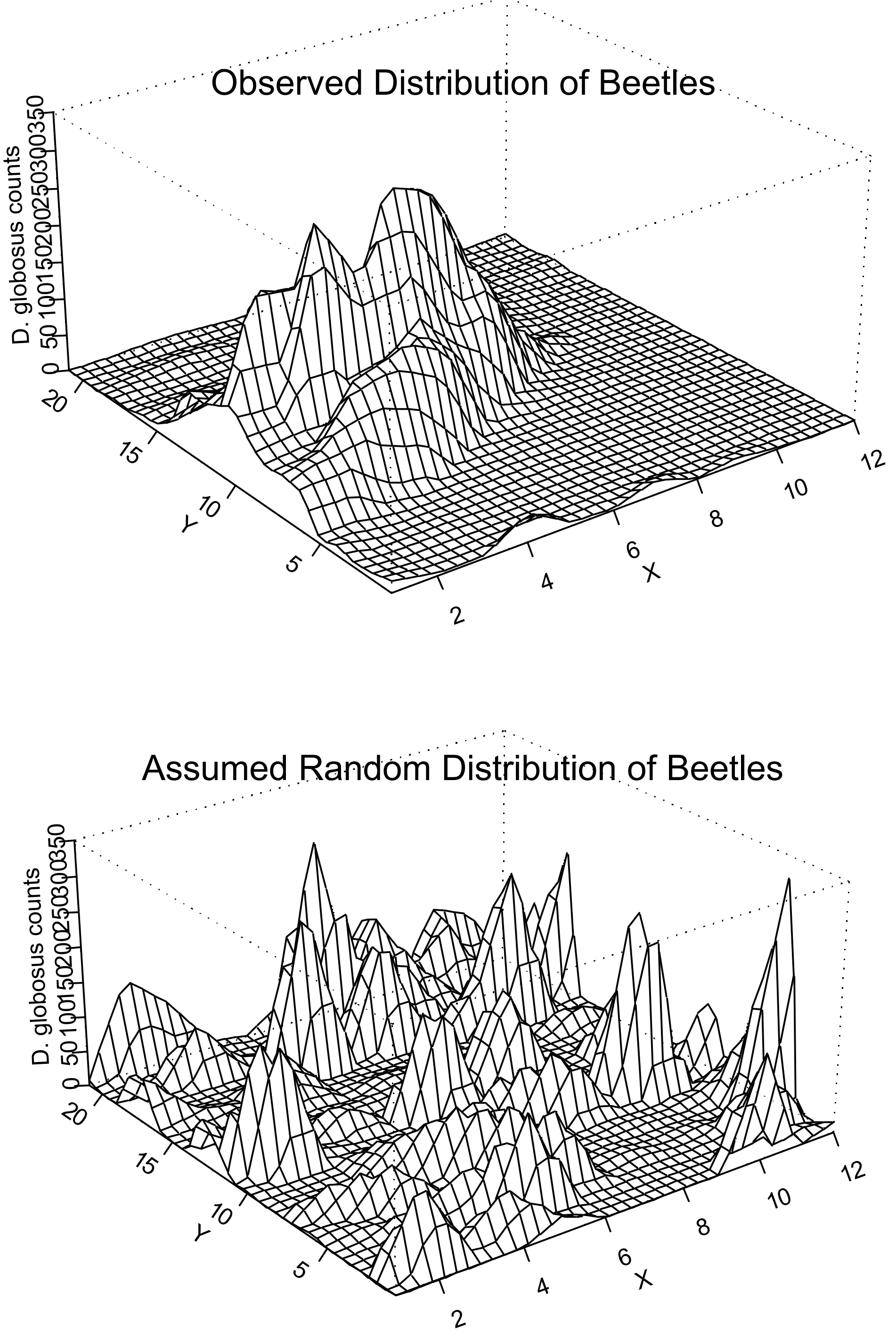
n=252 grid cells
Minimum = 0 beetles
Median = 2 beetles
Maximum =334 beetles
Mean=28.5 beetles
Std. Deviation =58.0

Given the large disparity between the mean and the median and between the mean and the standard deviation, we can conclude that the data are severely skewed and thus would require some sort of normalizing transformation before any of the parametric statistical tests described in Appendix D could be applied. However, all of these inferential procedures also require the data to be spatially and temporally independent. The independence assumptions cannot be assessed by inspection of sample descriptive statistics of the sort shown above.

Before considering the types of statistics that are appropriate for the evaluation of spatial autocorrelation, the nature of the problem needs to be illustrated. A 3-dimensional plot of the spatial distribution of the beetle counts on the sampling grid is shown in Figure 9A. The X- and Y-axes are the latitudinal and longitudinal dimensions of the sampling grid and are divided into grid cell units. The third axis provides a 3-dimensional surface representing the beetle counts. The graph reveals an area of intense population density near the center of the grid, with a complete absence of beetles elsewhere. In fact, 81 of the gird cells have counts of zero. Thus, if the beetle count in any given grid cell is known, there is a very high probability that adjacent grid cells will have similar numbers of beetles; i.e., the beetle counts of neighboring sampling units are highly correlated. This is equivalent to saying that the beetle counts are not independently distributed over the study area. For reference, Figure 9B depicts one of many possible configurations of beetle abundance that are randomly (i.e., Poisson) distributed in space.

The statistical procedures discussed in this appendix require spatially referenced sampling units to be distributed like those shown in Fig. 9B. When this is so, knowledge of the count in a specific grid cell does not provide any information on the counts in neighboring grid cells. By contrast, the closer any two grid cells in Fig. 9A are to one another, the more similar their beetle counts. Thus adjacent gird cells provide redundant information regarding the distribution and abundance of beetles in the study area. A sample of 30 grid cells that are far enough apart from one another to be uncorrelated will provide more information on the variation in the population distribution than will a sample of 30 that are close together. Likewise, any estimate of population mean or median abundance from a sample of uncorrelated grid cells will be unbiased whereas estimates from the correlated grid cells will be biased towards the particular subset of grid cells that are autocorrelated.

Fig. 9. Spatial distribution of beetles (*Dychirius globosus*) on a gridded study area. Grid cells are 40 m on a side. (A) Actual observed distribution in spring 1975 ($\mu = 28.5$, $\sigma = 58.0$) (B) Randomly reassigned distribution in spring 1975 ($\mu = 28.5$, $\sigma = 58.0$).



One obvious approach to diagnosing autocorrelation in a sample would be to estimate the degree to which neighboring sampling units are correlated or alternatively independent from one another. But first some criteria are needed as to what constitutes a neighbor. In fact the degree to which any two sampling units will be spatially correlated is a function of how far apart they are. Thus it is conventional to estimate a series of autocorrelation or variance values based on the distance between neighboring sampling points. These distances are called nearest neighbor distances. Figure 10 is a plot of the variance (vertical axis) among grid cells that are 1,2,3, ….12 grid cells apart. Spatial analysts call such nearest neighbor distances, **lags** (horizontal axis). In this case, a lag unit is equivalent to the length of the side of a grid cell; i.e., 40 meters. The curve labeled "Observed" is a plot of the nearest neighbor variances of grid cells from the study area (Fig. 9A) while the nearest neighbor variances of the "Random" curve were computed from the simulated independent data shown in Fig. 9B. The horizontal reference line marks the value of the overall sample variance (3368.6).

The graph, called a **variogram**, confirms that the variability in the beetle counts from the study area tends to increase with the distance of a sampling unit from it s nearest neighbor. Conversely the closer two grid cells are to one another, the more similar are their beetle abundances, hence the smaller their neighbor-neighbor variability. By comparison, the variability among grid cells from the randomly distributed population does not vary significantly with lag distance; moreover, it never deviates much from the overall sample variance, a pattern characteristic of spatially independent data. The sharp inverted J-shaped pattern of the variogram of the observed data is emblematic of a sample with serious autocorrelation.

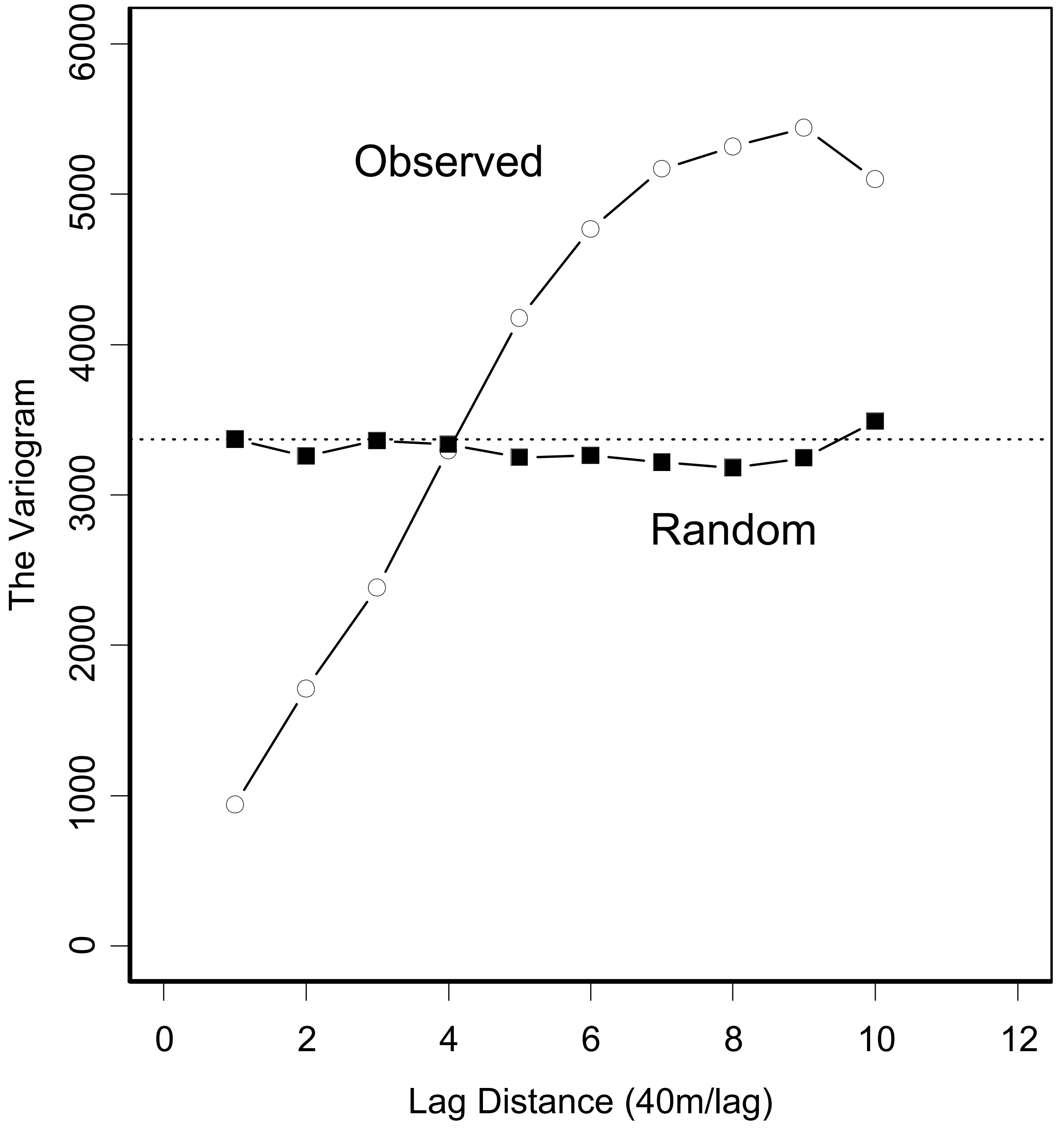
Two important conclusions can be drawn from these variograms: (1) Because the variance is an increasing function of lag-distance, the beetle count data are autocorrelated and (2) the negative effects of the autocorrelation do not extend beyond four lag-units (160m). The last conclusion is based on the lag distance associated with the intersection of the plot of the observed variogram with the overall sample variance reference line, which of course is the variance one would expect from a random independent sample.

Once spatial autocorrelation of the magnitude observed for the beetle data has been diagnosed, statistical tests and confidence intervals of the sort described in this appendix are not valid. Although there are number of methods available for the analysis of spatially and/or temporally correlated data, they are quite complex mathematically and require the assistance of an experienced spatial data analyst, using specialized software (e.g., ArcView Spatial Analyst or SPLUS SpatialStats). For water quality attainment studies, the best strategy will almost always be to construct environmental sampling designs in such a way that spatial correlation is minimized. This can be accomplished most easily be variogram analysis of appropriate pilot study data. For example, Hengeveld could have used the above results to design a fall sampling protocol wherein grid cells could not be selected if they were within 160 m of a previously selected grid cell.

C.2.5 EDA Example 4: Assessing Temporal Independence

The problems associated with temporal autocorrelation are similar to those of spatial autocorrelation, with the notable exception they occur in only one dimension (i.e., time). Just as

Fig. 10. Variograms comparing changes in the variance of beetle abundance with distance between grid cells for the observed field data (open circles) and the randomly reassigned counts (solid squares). The broken horizontal reference line marks the overall sample variance (58^2).



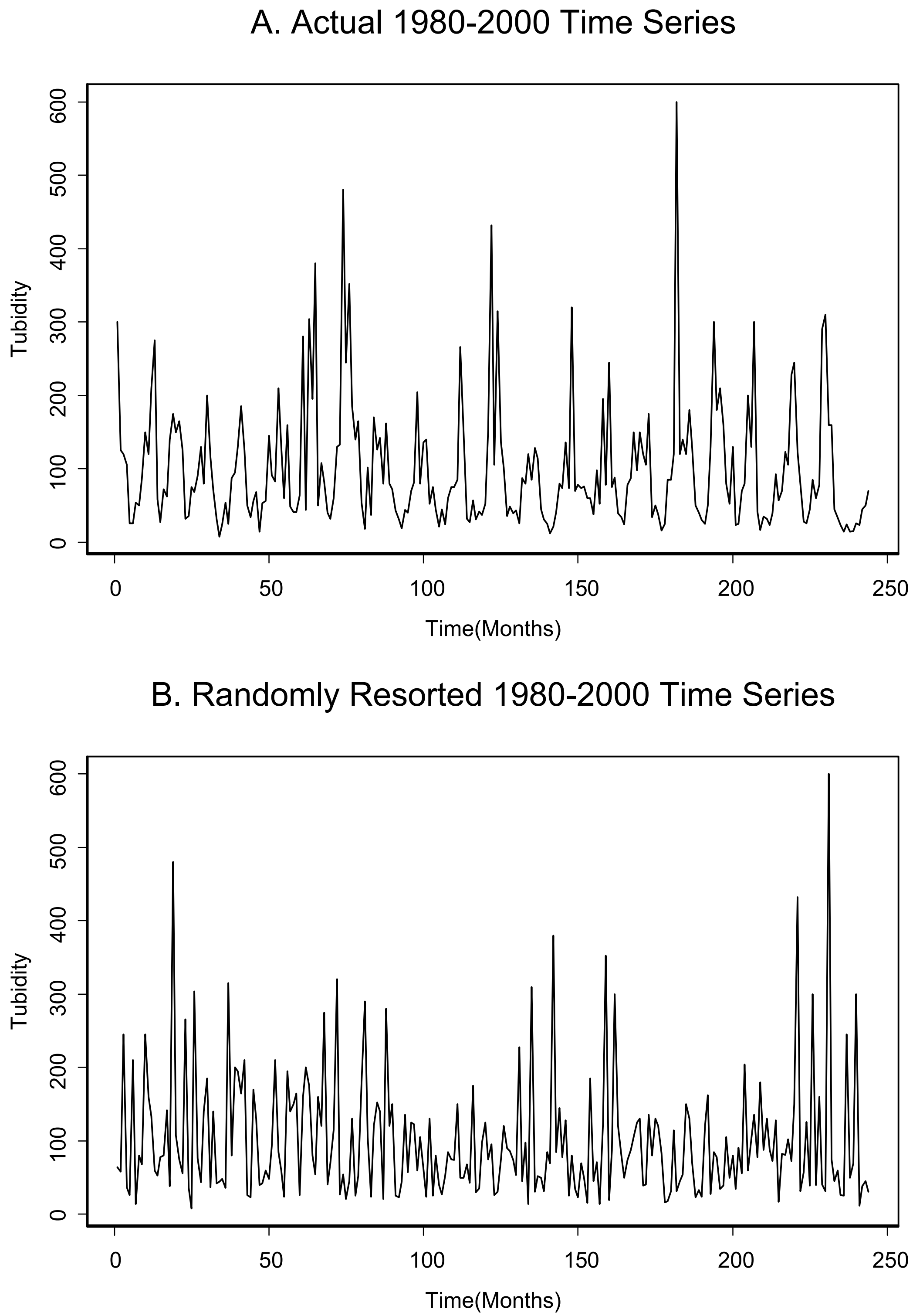
units sampled close together in space tend to be alike, so do units that are sampled close together in time. Not surprisingly then, the graphical techniques that are employed to diagnose temporal autocorrelation are similar to those used to assess spatial autocorrelation. As pointed out in Example 4, increasing correlation implies decreasing variability and *vice versa*. However, while it is customary to examine spatial autocorrelation by plotting the complementary variance relationships in variograms, time series analysts traditionally plot temporal autocorrelations directly in graphs called **correlograms**. In this example, correlogram analysis of patterns of temporal autocorrelation will be illustrated for the monthly turbidity data that were analyzed in Example 1 (Table 3).

The monthly turbidity time series is plotted in Fig. 11A. The series starts in January 1980 and terminates in April 2000. The plot reveals December-January spikes and corresponding May-July troughs that reoccur annually. This pattern suggests that seasonal factors (e.g., precipitation) may be influencing turbidity readings in the Mermentau River. Within any 1-year cycle it appears that monthly values tend to increase just before the spikes and decrease just after them. This suggests that measurements adjacent to the midwinter spike tend to be more alike than non-adjacent months. Thus the time series provides strong visual evidence of both seasonality and autocorrelation. For reference, a randomly ordered time series was generated by randomly reassigning month/year dates to the 244 monthly turbidity values, resorting and finally replotting them (Fig. 11B). The result is a chaotic series of random spikes and troughs with no apparent annual cycling. The confidence intervals and statistical tests described in the appendix assume a pattern similar to Fig. 11B and perform poorly for data with patterns like Fig. 11A.

Figures 12A and 12B are the correlograms which were computed for the actual and random time series in Figs. 11A and 11B, respectively. Each correlogram is a plot of the correlation (vertical bars) among monthly turbidity values that are 1,2,3, ….60 months apart. As in the variograms, these interval distances (horizontal axis values) are called lags. The first lag is zero; its associated autocorrelation bar is a measure of the correlation between each monthly turbidity value and itself, which of course is 1.0. The vertical bar above second lag is a measure of the correlation between each monthly turbidity value and that of the next month. Of particular interest are the correlations between turbidity values that were taken twelve months apart. These are the January 1980-January 1981, February 1980-Feberuary 1981, December 1980-December 1981 correlations, as well as the corresponding monthly correlations across any pair of successive years. Likewise, lag 24 correlations are the correlations between monthly turbidity values that are taken 2 years apart, lag 36 correlations are the correlations between monthly turbidity values that are taken 3 years apart, and so on. The two broken lines are the 95% confidence limits about the zero autocorrelation value. Thus any vertical bar that does not extend beyond the broken lines indicates an autocorrelation that is not significantly different from zero; i.e., such values denote temporal independence.

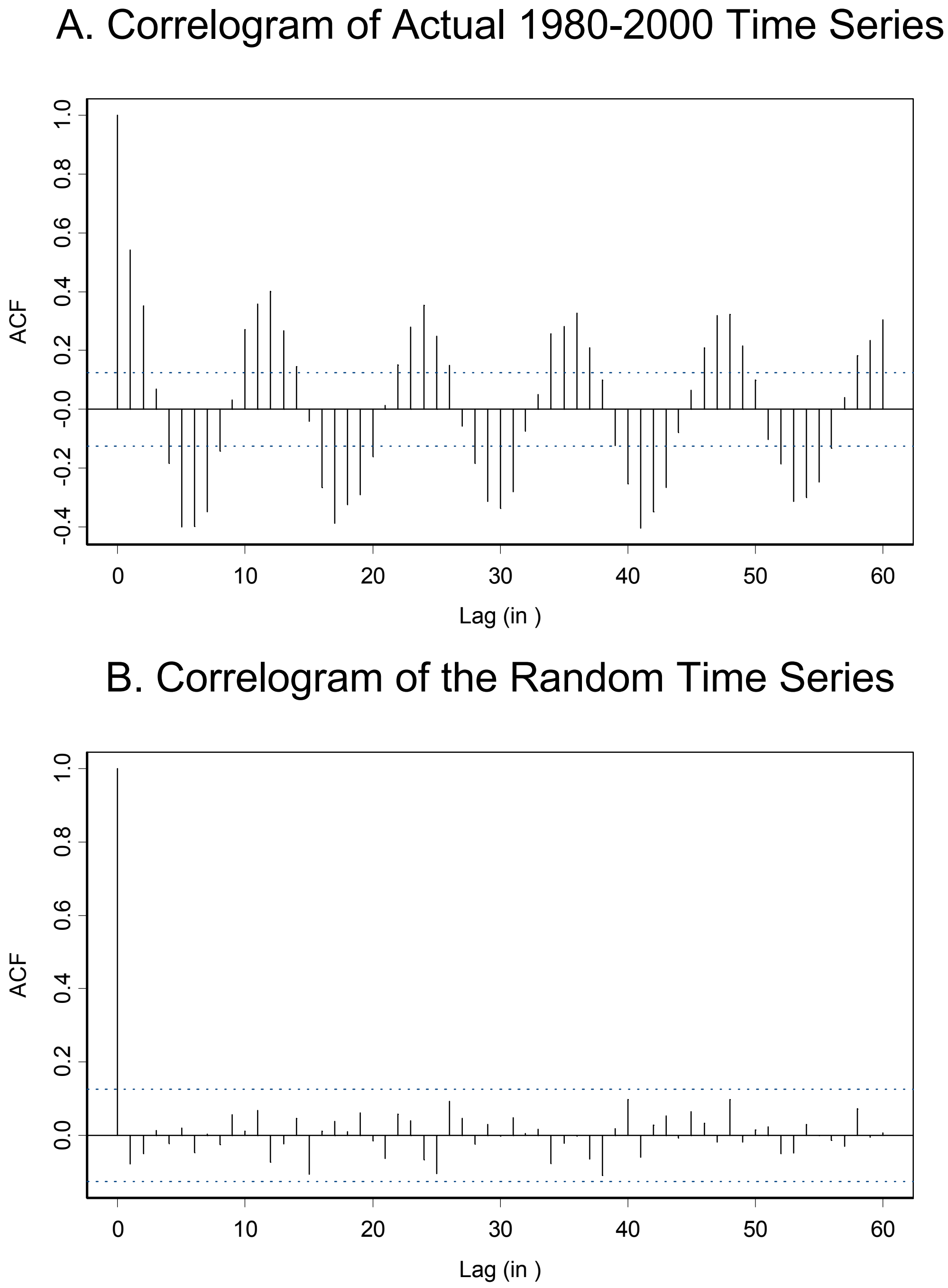
Several relationships are clarified by the two correlograms in Fig. 12. First, while there is no significant correlation beyond lag zero for the randomly resorted data (Fig. 12B), there is a consistent, recurring pattern of significant, alternately positive and negative, autocorrelations in the actual field data (Fig. 11A). The pattern in the field data repeats itself regularly at twelve-month intervals. Within a given year, the alternating positive and negative values simply reflect the up and down seasonal oscillation of the turbidity values in the Mermentau River. Significant

Fig. 11. Time series plots of 244 monthly turbidity measurements from the Mermantaut River, January 1980- April 2000. (A) The actual field data. (B) The field data randomly reordered in time.


A. Actual 1980-2000 Time Series

B. Randomly Resorted 1980-2000 Time Series

Fig. 12. Correlograms showing changes in the autocorrelation of turbidity measurements with vs. length of the time (lags) between measurements for the observed field data (A) and the randomly reordered data (B). Broken horizontal reference lines are the 95% confidence limits around zero autocorrelation; bars within them indicate temporal independence at those lags.



A. Correlogram of Actual 1980-2000 Time Series

B. Correlogram of the Random Time Series positive autocorrelations occur for lags (i.e., intervals) of 1, 10, 11 and 12 months with the highest correlations associated with the twelve-month lags. Significant negative autocorrelations occur for lags of 4, 5, and 6 and 7 months with the largest negative correlations associated with turbidity readings that were taken 5 or 6 months apart. Near-zero autocorrelations occur among measurements taken 3, 8, or 9 months apart.

The correlogram of the field data (Fig. 12A) suggests that analysis and/or data collection should be confined to data that have been collected at 3, 8 or 9-month intervals. This poses some serious problems to regulators who want to base decisions on three-year evaluation periods.

Since shorter sampling intervals yield larger sample sizes, the optimal choice is a 3-month interval, but this only yields twelve samples per three-year period. As will be discussed in the next section, small samples generate large decision error rates for any of the statistical tests and confidence intervals discussed in this appendix. On the other hand, autocorrelated data tend to produce biased results and hence erroneous decisions. So, what should be done? There are several alternative solutions to this common environmental data problem; one of the simpler solutions will be demonstrated in Appendix D (Section D.2).

C.3 Data Quality Assessment: Hypothesis-testing and Estimation

C.3.0 Use Of EDA and DQOs to Select an Appropriate Hypothesis-Test or Estimator

A rigorous DQO process should have included the specification of the statistical tests that are appropriate to supporting the water quality attainment decisions that motivated the study. In Step 3 of the DQA, the task of the analyst is to review, revise, and if necessary replace these tests in light of the results of the EDA. Thus this step is essentially a more informed reenactment of DQO steps 5 and 6. In Sections C.1.1-C.1.5., these two steps were discussed from the perspective of confidence interval construction. Although there is a strong natural connection between confidence intervals and statistical hypothesis tests, there are some important differences. In Sections C.3.1 – C.3.3 we will explore these differences and in Appendix D provide guidance on how to use parametric and nonparametric hypothesis tests to compare a sample mean or proportion to an environmental standard in a water quality attainment context.

C.3.1 Hypothesis-Testing Basics

Statistical hypothesis testing is broadly applicable to many situations beyond what are described here. In fact, the present application is an example of the simplest of all statistical testing scenarios: the so-called one-sample question. **One-sample tests** are typically used to evaluate hypotheses about whether some sample statistic (e.g., a mean, median or percentile) is equal to or exceeds a threshold such as a water quality standard. Tests useful for comparing two samples (e.g., a sample from an ambient or treated population to one from a reference or a control population) and tests for simultaneous comparison of samples from several populations (e.g., ANOVA and linear regression) are covered in *Guidance for Data Quality Assessment EPA QA/G-9,* in *Biological Criteria: Technical guidance for Survey Design and Statistical Evaluation of Biosurvey Data* (EPA 822-B-97-002*),* and in many standard statistical texts (e.g., Steel et al. 1996; Millard and Neerchal 2000). All of the tests described in this document can be implemented with standard commercial software (e.g., SAS and SPLUS); most of the additional tests described in *Guidance for Data Quality Assessment EPA QA/G-9,* can be carried out with SPLUS EnvironmentalStats.

Hypothesis testing is motivated by the need, during the decision-making process, to account for uncertainty in data collected by a **probability-based sampling design.** There are several possible sources of uncertainty including:

1. Sampling variation specific to the design employed to collect the data
2. Intrinsic natural (e.g., genetic or behavioral) variation among population members
3. Temporal and/or spatial variation
4. Measurement and/or laboratory error
5. Model misspecification error (e.g., in Monte Carlo risk assessments)

In a typical water quality attainment study, we usually evaluate either of two pairs of one-sided null and alternative hypotheses. The scientific objective is to compare an unknown population parameter against a specified value that prior studies have indicated is a threshold for desirable/undesirable environmental outcomes. Depending on the situation the hypothesis pair may be either,

$$H_0 : q \leq q_0 \quad vs. \quad H_a : q > q_0 \tag{2}$$

or,

$$H_0 : q \geq q_0 \quad vs. \quad H_a : q < q_0 \tag{3}$$

where θ = the population parameter of interest
θ_0 = the fixed criterion value that the population parameter is compared against

Equation 2 evaluates an upper one-sided alternative hypothesis, while Eq. 3 evaluates the complementary lower one-sided alternative. As an example of the application of the former, we may want to evaluate the null hypothesis that the mean sediment concentration of pesticide X is less than or equal to 20 µg/Kg vs. the alternative that the sediment concentration exceeds 20 µg/Kg.

Having selected the population parameters of interest (e.g., the mean concentration of a pollutant) and chosen null and alternative hypotheses appropriate to the water quality attainment decision under consideration, the next step is to choose a statistical test which can be used to determine which hypothesis (i.e., H_0 or H_a) is better supported by the sample data. Statistical tests are simple mathematical models that predict the distributions of test statistics when the null hypothesis is true. These are the distributions one would expect to obtain from conducting thousands of surveys or experiments and plotting the frequencies of the computed test-statistics. These distributions, called **sampling distributions**, reflect the uncertainty in the sample estimates of the values of the test statistic.

Statistical tests are commonly named for their sampling distributions (e.g., t-test, chi-squared test). The test statistics themselves are usually simple algebraic functions of the sample statistics. For example, the test statistic for the test against either one-sided alternative for comparing the mean of a continuous random variable, such as the sediment concentration of pesticide X, to a fixed value such as a 20 µg/Kg environmental standard is a function of the sample size (n), mean ($\bar{x}$) and variance (s_x^2):

$$\frac{\bar{x}-20}{\sqrt{s_x^2/n}} \tag{4}$$

Statistical theory insures that when certain assumptions hold and the null hypothesis is true, the sampling distribution of the test-statistic shown in Eq. 4 will be a t-distribution with n-1 degrees of freedom (df). Thus the associated statistical test is called a t-test. There is actually an entire family of t-distributions, each with a different df.

The sampling distributions of the test statistics should not be confused with the population distributions from which the samples have been collected. Whereas the latter (e.g., Fig. 5a) are the natural distributions of the animals or water quality factors under study, the former are statistical models of the behavior of statistics calculated from samples of the natural populations. Some statistical tests, called **parametric tests** (e.g. t-tests) require that the natural populations be normally distributed, while others, called **nonparametric tests** (e.g., chi-square tests), make no assumptions about the distribution of the natural populations.

All statistical hypothesis tests are mathematical models. In the case of t-tests, chi-square tests and F-tests (i.e., tests associated with ANOVA and regression models), the distribution of the test statistic computed from the sample data is modeled as (respectively) a t-distribution, a chi-square distribution or an F-distribution. Like all models, the validity of the predicted distributions depends on assumptions made about the underlying processes that are being modeled. In the case of one-sample t-tests for the population mean, the following assumptions are made:

1. The variable being analyzed is a continuous random variable that is normally distributed in its target population.
2. The sampling units used to compute the sample mean from which the t-statistic (Eq. 4) was computed were independently distributed in the target area (i.e., there is no temporal or spatial autocorrelation among the sampling units).
3. There was no systematic error associated with measuring the response on the sampling units (e.g., pH meters and/or laboratory assays were correctly calibrated and applied).
4. The null hypothesis is true.

The distribution of the test statistic under the null hypothesis is the basis for determining whether the data support the null hypothesis or the alternative. For example if we have a sample of 30 sediment sampling units, the expected distribution of t-statistics under the null hypothesis is a t-distribution with df=29. Ninety-five percent of the t-statistic values in such a distribution are less than 1.699. Thus if the t-statistic value computed from our sample has a value of (say) 8, then we conclude that there is less than a 5% probability that our sample came from such a t-distribution. This suggests that one or more of the above four following assumptions is *not* true. If we have previously verified the first three assumptions (e.g., by application of the appropriate DQOs and EDA methods in Sections C.2.1- C.2.5), we can conclude that our sample *does not* support the null hypothesis. If we have not verified the assumptions, we cannot draw any conclusions from the t-test. Rejection of the null hypothesis provides evidence in favor of the alternative that the sediment concentration of pesticide X exceeds the environmental standard of 20 μg/Kg.

The probability used as the cutoff for accepting or rejecting the null hypothesis is called the **significance level**. By declaring a significance level of 5%, we are saying that even though there is a 5% probability that a t-statistic ≥ 1.699 could have come from the null t-distribution, this probability is so small that we don't think it is reasonable to believe that the data actually came from a target area whose pesticide X concentration was ≤ 20 μg/Kg. On the other hand, there *is* a 5% chance that it could have. Thus the significance level is just the Type I error rate (α) that the investigator decided in DQO step 6 he would be willing to live with.

C.3.2 Types I and II Error Rates and Statistical Power

In this section graphs of the distribution of the test statistic when the null hypothesis is true will be compared to its expected distribution when it is false. The graphs will be used to clarify the relationships between the two types of decision error and the statistical power of the hypothesis test. The important effects of three attributes of the sample on the power of the t-test also will be explored and illustrated through the use of graphs. These factors are: the sample size, the sample variance, and the ± difference between the sample estimate of the population parameter (e.g., the pollutant mean concentration) and the criterion value that is the basis for an attainment/impaired classification. As will be seen, the observed power of any statistical test (*not* just the t-test) is the result of a complex interaction of these three factors and the investigator's choice of the null and alternative hypotheses.

We will begin by examining the hypothesized distributions of pesticide X in the sediments of the target area and the distributions of the corresponding t-test statistic under both the null and alternative hypotheses. Recall that alternative hypothesis is open ended; i.e., it states only that the true sediment mean concentration is greater than the criterion value of 20 μg/Kg. Thus any value > 20 will be consistent with H_a. Figure 13 shows the hypothesized distributions of the sampling unit concentrations of pesticide X as they are expected occur in the sediments of the target area if H_0 is true (normal: μ= 20 and σ = 1) and a possible alternative distribution if H_a is true (normal: μ= 21 and σ = 1). Using the sample mean and standard deviation, a t-statistic for testing the null vs. the alternative hypothesis can be calculated with Eq. 4.

The influence of sample size on the expected distributions of this t-statistic under the null and alternative hypotheses are shown in Figs. 14a and 14b . The distribution of a t-statistic under the alternative hypothesis is called the **noncentral t-distribution**. Based on a sample size of 30 sampling units (Fig. 14b), the two distributions are quite distinct with only a sliver of overlap. In contrast, the distributions of test statistics computed from the much smaller sample size of 10 (Fig. 14a), overlap considerably. By analogy, we can think of the distributions as distant mountains viewed through a telescope. While the two peaks are distinct when observed through a powerful telescope (Fig. 14b), they appear to merge when seen through a weaker telescope (Fig. 14a). The heavy vertical line in Fig. 14a marks the point at which the t-statistic values of the alternative distribution are less than the 95th percentile of the null t-distribution. The noncentral t-values that lie to the left of this line represent t-test outcomes that are erroneously regarded as evidence that the data come from the null distribution. This proportion of the alternative distribution is the Type II error probability (β); it is the part of the "alternative mountain" that can't be distinguished from the "null mountain." The complementary proportion (1-β) of the alternative distribution to the right of the 95th percentile of the null, is the part of the alternative distribution that is *correctly* distinguished from the null distribution. The larger this proportion, the greater the resolution of the test. Thus 1-β is a measure of the **power** of the test to correctly identify t-statistic values that support the alternative hypothesis. It is clear from Fig. 14a why, for fixed sample size and variance, decreasing the α value increases the β value; decreasing α moves the heavy vertical line to the right, causing more of the alternative to "merge" with the null distribution.

The influence of the sample standard deviation (σ) on the power of the t-test is illustrated in Figures 15a and 15B. All the distributions shown in Fig. 15 are based on sample sizes of 30 (df=29) from populations with the same standard deviation of 15. Like the distributions in Fig. 14b, the null and the noncentral t-distributions in Fig. 15a have means 20 and 21, respectively. Thus the only difference between the Fig. 14b and 15a distributions is that the former have σ=1 while that latter have σ =15. However, this 15-fold increase in variance results in a decline

Fig. 13. Distributions of two populations of sediment sampling units, both of which are normal with standard deviations of one, but with different means (solid line = mean of 20; broken line = mean of 21). The null hypothesis states that the former is the true population distribution, while the latter is one of several possible population distributions that could occur if the alternative hypothesis is true.

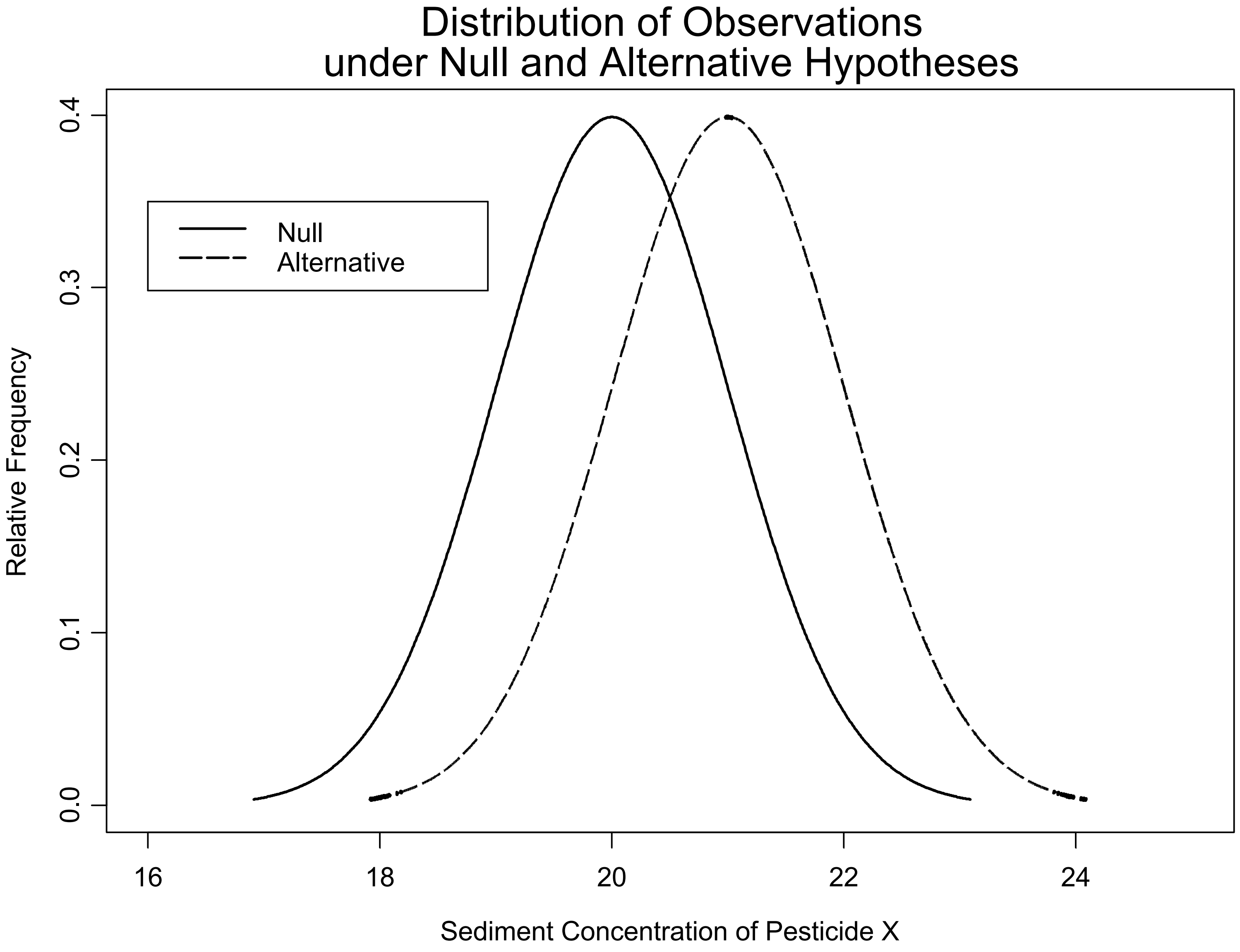


Fig. 14. Sampling distributions of t-statistics with both their population standard deviations and the effect size equal to 1 when the null hypothesis is true (solid line) vs. when the alternative hypothesis is true (broken line). (A) compares the two distributions when the sample size is 10; (B) compares them when the sample size is 30.

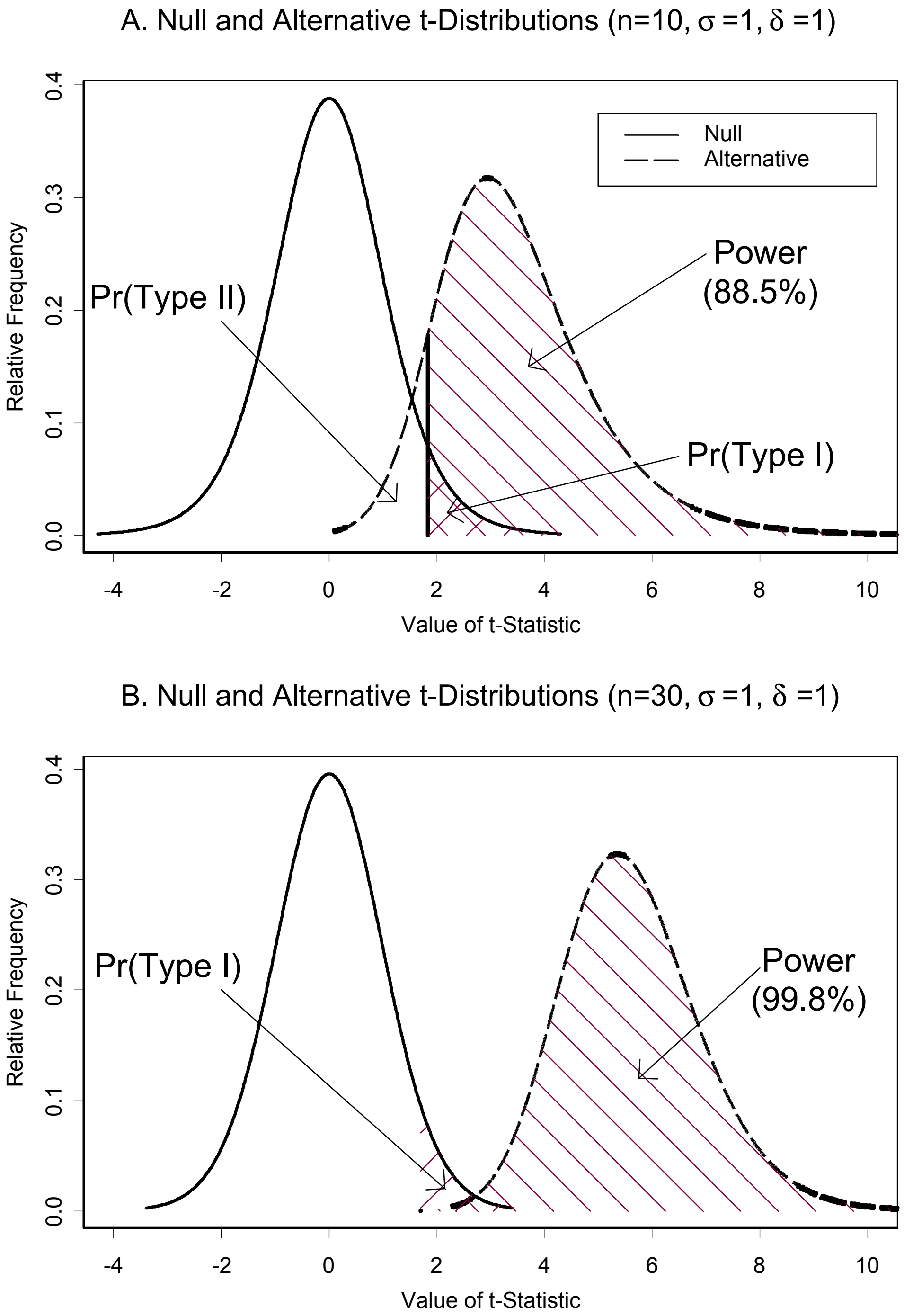
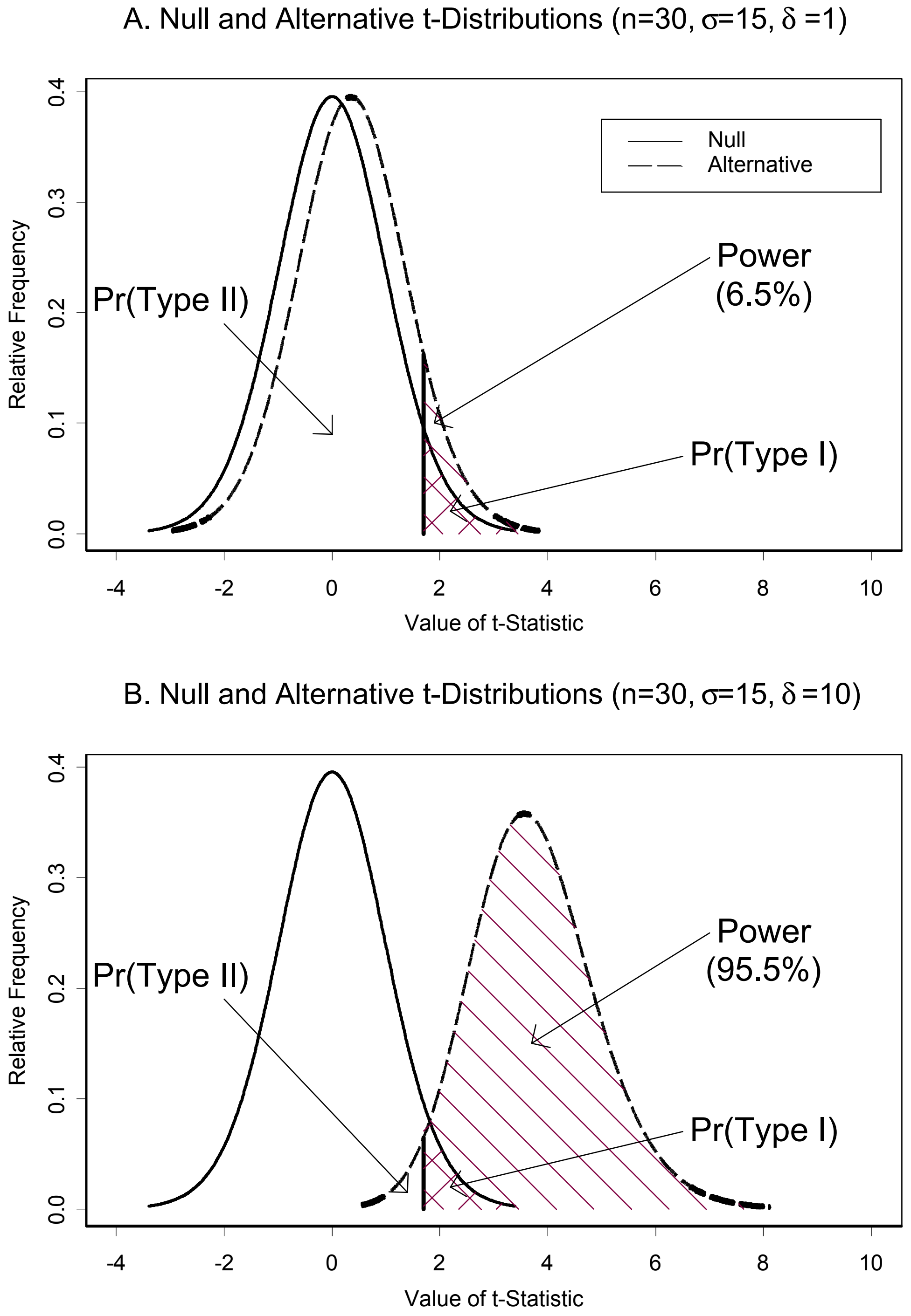


Fig. 15. Sampling distributions of t-statistics with population standard deviations of 15, and sample sizes of 30 when the null hypothesis is true (solid line) vs. when the alternative hypothesis is true (broken line). (A) compares the two distributions when the effect size is 1; (B) compares them when the effect size is 10.



A. Null and Alternative t-Distributions (n=30, σ=15, δ =1)

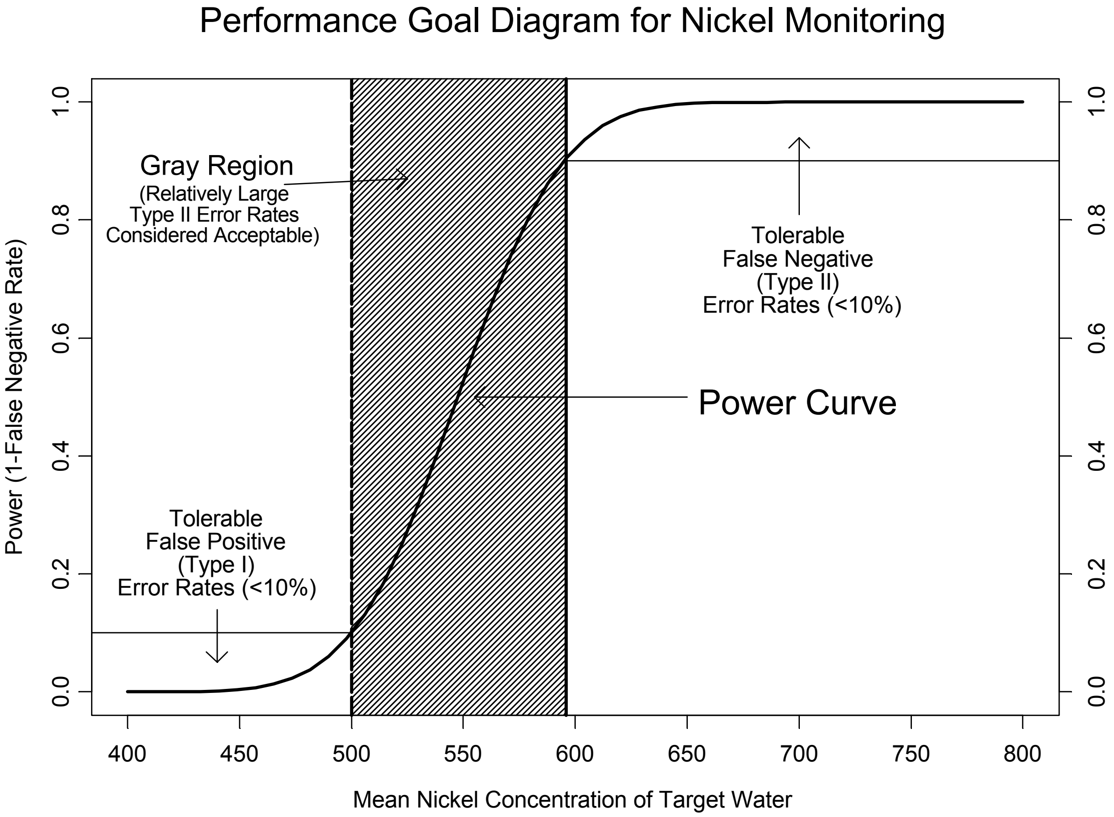
B. Null and Alternative t-Distributions (n=30, σ=15, δ =10) in power from 99.8% to only 6.5% for the fixed sample size of 30. As illustrated in Fig. 15a, the null and the alternative distributions occupy nearly identical locations with only a small proportion (6.5%) of the alternative distribution to the right of the 95th percentile of the null distribution. This illustrates the general principle that an increase in the variance of a sample will cause a decrease in the power of an statistical tests based on that sample.

The effect of the size of the difference between the sample mean and the criterion value on statistical power can be seen by comparing Fig. 15a with Fig. 15b1. In Fig. 15b, the noncentral t-distribution has a mean that is 10 units larger (i.e., 30) than the null distribution. This has the effect of shifting its location 10 units to the right of the null, simultaneously reducing the Type II error (i.e., the overlap with the null) and increasing the power to 95.5%. Returning to our telescope analogy, we note that if one object is located directly in front of the other, it is nearly impossible to distinguish them. But if one of the objects is shifted away from the other, it becomes discernible.

In the language of experimental design, the observed difference between a mean (or median) and the criterion value to which it is being compared is called the **effect size**. Figures 15a and 15b illustrate the general case: the smaller the effect size (usually symbolized by δ), the more difficult it is to for a statistical test to distinguish the sample mean (or median) from the criterion value; i.e., the power is low. For fixed δ and population σ, the only way to increase the power to a point where the difference becomes "visible" (i.e., H_0 is rejected), is to increase the sample size. Thus whenever a statistical test will be used to compare a sample mean or median to a criterion value, it is important decide *a priori* how large the difference (i.e., δ) must be to be "ecologically significant". Once this is done, an appropriate minimum sample size can be estimated. Statistical significance for a study designed in this manner will be a reasonable indicator of ecological significance. If there is not a consensus on how large the effect size should be for it to be ecologically or toxicologically significant, it will be difficult (perhaps impossible) to use the results of the statistical test to support a WQS decision.

Figure 16, called a Decision Performance Goal Diagram, presents a graphic summary of the relationship of effect size to the false negative rate (i.e., Type II error rate) and its complement the statistical power associated with either a lower one-sided 90% confidence interval or a t-test with an upper one-sided alternative. In this example, the sample mean concentration of nickel is compared to a CMC 500 µg/L. The null hypothesis is that the concentration is $\leq$ 500 µg/L vs. the alternative that it is > 500 µg/L. The sample size is 30 one-liter sample units drawn from water body of interest and the standard deviation is 200. The vertical axis values are the power (i.e., $1-\beta$) and the horizontal axis displays the reasonable expected range of the nickel concentrations in the sample. The power curve traces the power against the one-sided alternative hypothesis. Note that when the sample means are low (<450 µg/L) the false positive (i.e., Type I) error rates and the power are near zero. This is at is should be; sample means in this range do not provide much basis for rejecting H_0. However as the sample means increase from 500 to 596, the statistical power increases rapidly. This "gray" area represents the range of values where inference is unreliable; it is bounded on the left by the criterion value and on the right by the criterion value + the effect size (i.e., δ=596-500 µg/L). The right bound of the gray area is the value at which the sample size of 30 provides sufficient power to insure rejection of H_0 in $\geq$ 90% of samples. Thus Fig. 16 suggests that given a sample size of 30 sampling units and an

Fig. 16. Decision performance goal diagram for a one-sample t-test of H_0: population mean ≤ 500 vs. H_a: population mean >500, when the standard deviation =250, the sample size=30 and the minimum detectable effect size=100.



effect size of 96 µg/L, the t-test will have both good power and low false positive error rates. Once the mean exceeds 596 µg/L, H_0 will be rejected with probability s ≥ 0.90 (i.e., false negative error rate < 0.10). From a confidence interval perspective, specifying a narrow half-width on the interval estimate is analogous to decreasing the minimum detectable effect size δ; either of these actions will decrease the width of the gray area by moving the right bound closer to the criterion value, thereby increasing the area with tolerable Type II error rates to the right of the gray zone. See Chapter 6 in "Guidance for the Data Quality Objectives Process" (EPA QA/G-4) for additional discussion of Decision Performance Goal Diagrams.

For a one-sample test, the effect size is essentially a detection limit. Although one can set a criterion such as 500 µg/L for a concentration, if one employs a one-sample t-test or one of its nonparametric counterparts, the result of the test will be depend on the whether the difference between the sample mean or median and 500 µg/L criterion is large enough to be statistically significant. The effect size tells us how much the sample mean must exceed the criterion value before the exceedance is detected with a probability of 1-α. Thus the effect size effectively resets the criterion value to a new value equal to the sum of the original value + the effect size (e.g., 500 + 96 µg/L). This adjustment is an acknowledgment of the effects of the uncertainty in the sample estimates on the Type II error rate. Perhaps, just as important, in setting of the minimum effect size, the investigator must recognize the cost constraints on the sample size, " In essence, the gray region [which is determined by δ] is one component of the quantitative decision performance criteria that is specifically used to limit impractical and infeasible numbers of sampling units" (GS-4, page 6-4).

In summary, the observed power of a statistical test is the result of a complex interaction of the sample size, the variance of the attribute being measured, the effect size, and the specified α-level. When each of the other three factors is fixed:

1. Decreasing the variance increases the power of a statistical test
2. Increasing the α-level (e.g., 0.05?   0.10) increases the power of a statistical test
3. Increasing the effect size (gray area in Fig. 16), increases the power of the test
4. Increasing the sample size increases the power of a statistical test

In general, it will not be advisable to base a WQA decision on the p-value from a hypothesis test unless the interactive effects of sample size, sample variance, effect size, and the specified tolerable Type I error rate (i.e., the α-level) on the power of the test have been carefully considered and are consistent with the DQOs and the objectives of the WQS.

C.3.3. Reversing the Null and the One-sided Alternative Hypotheses

Up to this point, our discussion of one-sided tests for comparing a population parameter θ (e.g., percent exceedance or mean pollutant concentration) to a regulatory standard k has focused on paired hypotheses of the form:

$$H_0 : q \le k \quad vs. \quad H_a : q > k \qquad (5)$$

These hypotheses lead us to define the Type I error rate as the probability of incorrectly classifying as "impaired", a body of water that attains the water quality standard. Similarly, the Type II error rate is defined as the probability of incorrectly classifying an impaired water as one that meets the standard. Because the Type I error rate (α) is fixed by the investigator or regulatory agency, its value is assured regardless of the sampling design implemented or the variability of the monitored water body. However, as shown in the preceding section and in Section D.5, the Type II error rate (β) depends on the complex interaction of the specified α, the sampling design, the sample size, the variance of the target water and on the resulting minimum detectable effect size (δ). This last point is especially vexing since δ effectively increases the minimally acceptable pollutant level from k to k+δ.

From the perspective of environmental protection, it can be argued that it is much more desirable to fix β rather than α. For example, by fixing β at 0.05, one can assure that there is a 95% probability that any water that exceeds the criterion value k will be correctly identified as impaired, regardless of the amount by which it exceeds k. However, as long as attainment of the water quality standard is the baseline condition, investigators and/or regulatory entities will not be able to control the Type II error rates.

A solution is available if the null and the alternative hypotheses are "flipped" like so:

$$H_0 : q \geq k \quad vs. \quad H_a : q < k \tag{6}$$

The flipping of the null and the alternative results in a concomitant exchange of the Types I and II errors. Thus, under the new set of hypotheses, α becomes the probability of incorrectly classifying, as attaining the standard, a water that is in fact impaired, while β becomes the probability that a water body that meets the attainment criterion is incorrectly classified as impaired. As in the former case (i.e., Eq. 5), the regulatory agency can fix α at 0.05 or any other rate of its choosing, thereby insuring that all but 5% of impaired waters are correctly classified, regulated and/or remediated.

There are three additional benefits to flipping the hypotheses:

1. Power now becomes a compelling concern of the monitoring entity (e.g., the polluter). This will provide a powerful incentive for it to devise the best possible sampling designs employing the largest affordable sample sizes, in an attempt to prevent the water(s) from being incorrectly listed. Consequently, the general quality of WQA studies would likely be improved greatly. The currently implemented testing scenario (i.e., Eq. 5) actually supplies a disincentive for doing this.

2. The new approach (Eq. 6) avoids the inevitable problem of specifying an effect size (δ) that increases the desired criterion value (k) to a new, less protective, threshold (i.e., k+δ; see also Fig. 16).

3. The new approach provides a financial reward to a monitoring entity that has kept pollution far below the criterion, since a much smaller n is required to protect against a Type II error

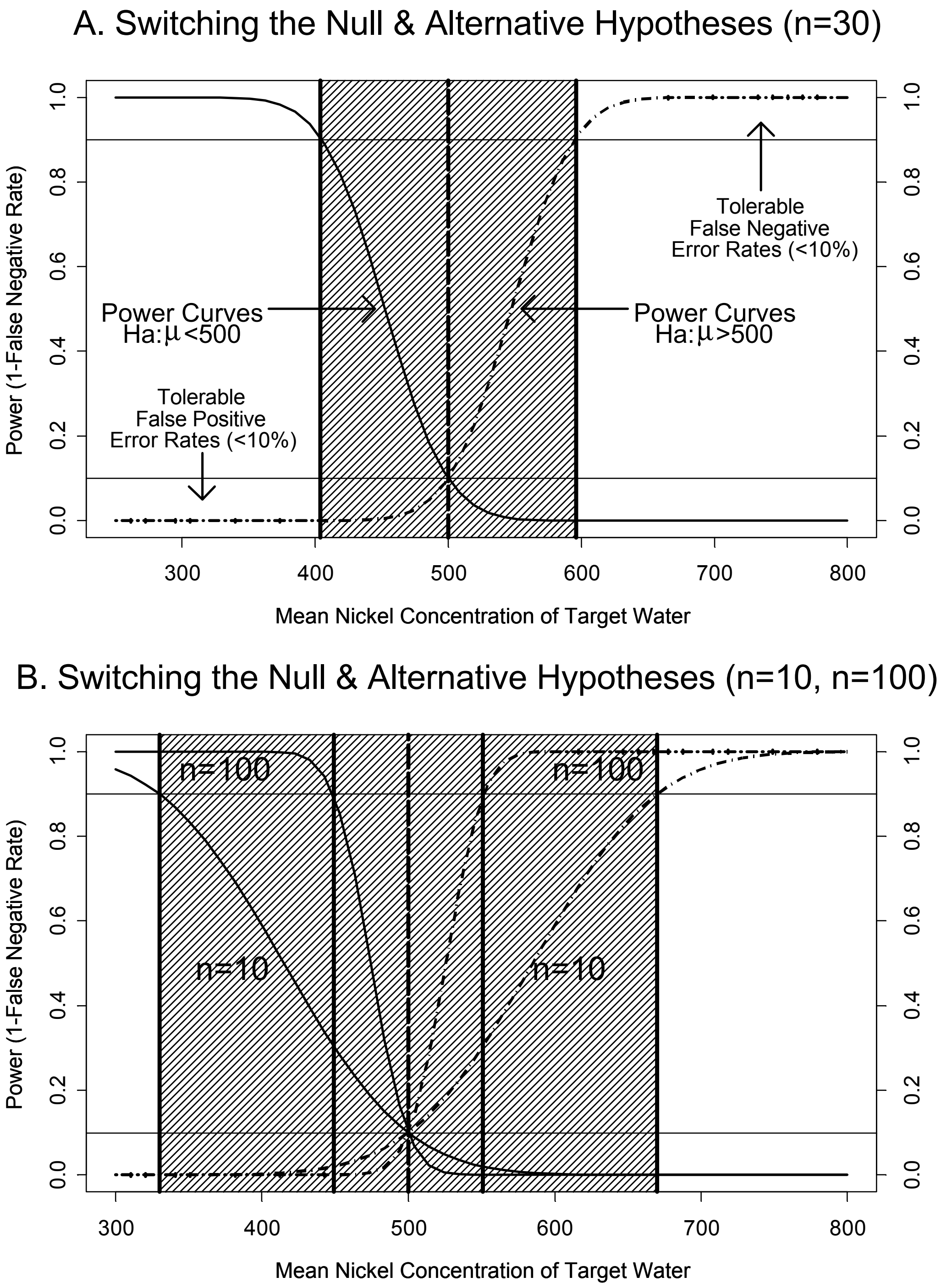
(i.e., to have high power), the farther the sample mean is below the criterion. Conversely, monitoring entities that have reason to believe the mean of the water under study is close to the standard will need to employ a much larger sample size to ensure a comparable level of power.

These relationships are illustrated in Figs 17a and 17b, which refer to the nickel monitoring problem described in Section C.3.2 and in Fig. 16. Figure 17a illustrates two power curves and their associated "gray regions", based on a sample size of 30 aliquots and a standard deviation of 200 µg/L. The broken line is identical to the power curve in Fig. 16; it corresponds to the null hypothesis that the mean nickel concentration in the target water is ≤ 500 µg/L vs. the alternative that the mean concentration is greater than 500 µg/l. The solid line is the power curve associated with the "flipped" hypotheses wherein the null becomes H_0: $\mu \geq 500$ vs. the alternative, H_a $\mu < 500$. A pair of horizontal reference lines, one near the top and one near the bottom of the graph, mark respectively the zones within which the false negative error rates (i.e., Type II) and false positive error rates (i.e., Type I) are < 10%. Because the population standard deviation is 200 and we have fixed the Type I error rate at 10% and the sample size at 30 aliquots, the minimum detectable difference (δ) between the sample mean and the criterion value (k=500) is 96 µg/L. This is true for both pairs of hypotheses; thus the gray region associated with H_a $\mu > 500$ extends from 500 µg/L to 596 µg/L while the gray region associated with H_a $\mu < 500$ extends 96 units in the opposite direction (i.e., 404 µg/L - 500 µg/L). The two gray regions appear respectively to the right and left of the vertical dashed line (at k=500) in Fig. 17 a.

The interests of the regulating agency focus on controlling errors to the right of the dashed 500 µg/L vertical reference line. As pointed out above, the regulating agency can specify and control the false positive error rate. When the alternative hypothesis is H_a $\mu < 500$, the area controlled (i.e., the area below the lower horizontal reference line) will be on the right side of 500 µg/L, whereas the area that is subject to the effects of the sample size, the variability, and the minimum detectable difference (i.e., the area above the upper horizontal reference line) will be on the left side. When the specified alternative is H_a $\mu < 500$, Fig. 17 a shows that whenever the sample mean concentration of nickel exceeds the 500µg/L criterion value, the probability that the associated water will be *correctly* classified as impaired *will always be at least 90%*, (i.e., the solid curve is always below the 10% false positive reference line for nickel concentrations > 500µg/L). The left-side gray region (i.e., 404-500 µg/L) corresponds to the range of sample means that have a fairly high probability of being incorrectly classified as exceedant (i.e., the false negative rate associate with H_a $\mu < 500$), even though the observed sample mean is less than the 500 µg/L criterion. It is to the advantage of the monitoring entity that the width this region be shrunk towards the criterion value. As explained in Section C.3.2, this is usually accomplished through optimizing the sampling design and increasing the sample size.

Figure 17b further illustrates the advantage to the monitoring entity of increasing the sample size when the mean concentration is believed to close to the criterion value; i.e., for situations in which there is a reasonable probability of exceedance. The solid lines are the power curves associated with H_a $\mu < 500$ when the sample sizes (n) are 10 or 100 aliquots. Similarly, the broken lines are the power curves for the complementary alternative, H_a $\mu > 500$ for n=10 or n=100. The inner pair of vertical reference lines embrace the gray regions associated with

Fig. 17. Decision performance goal diagrams contrasting power curves for studies employing one-sample t-tests with H_a μ>500 (solid sigmoid curves) vs. those with H_a μ<500 (dashed sigmoid curves). Panel A is based on a sample size (n) of 30; Panel B compares power curves based on n=10 to those based on n=100. See discussion, pages 25-26.



A. Switching the Null & Alternative Hypotheses (n=30)

B. Switching the Null & Alternative Hypotheses (n=10, n=100) n=100, while the outer pair bound the gray regions associated with n=10. Note that the solid vertical reference lines intercept the upper horizontal reference line at the points where each power curve crosses the 10% false negative error rate boundary and that regardless of which sample size is chosen, the false positive rates associated with H_a $\mu<500$ will always be less than 10% (lower horizontal reference line). A monitoring entity that is confident that the mean concentrations of nickel in the target water are lower than 300 μg/L will be satisfied with a sample size of 10 aliquots. However, if it suspects that mean levels are more likely in the vicinity of 450 μg/L, it will probably want to invest in a sample size of at least 100 because of the relatively high probability of erroneously listing the water if a smaller sample (e.g., 10 aliquots) is used. The choice of n=10 or n=100 will not be a matter of concern to the regulating agency since, in either case, the environment will be protected against falsely declaring attainment of the criterion.

Although it might appear that the balanced α-β approach described in Section D.5 would not be effected by reversing the null and alternative hypotheses, it fact there would likely be beneficial consequences because of the interest, on the part of the monitoring entities, in keeping β as low as possible. This would make it easier for the regulatory agency to obtain broad support for lowering the values of *both* α and β, thus improving the overall quality of WQ assessments.

As is the case for the currently implemented approach, one can base a WQA decision associated with the reversed hypotheses (Eq. 6) on 1-sided 100×(1-α)% confidence intervals. There are two possible 1-sided confidence intervals for the population mean concentration μ of a pollutant, the lower 1-sided 100×(1-α)% confidence interval,

$$\left[\bar{Y} - t_{n-1,1-a} \times \frac{s_y}{\sqrt{n}},\ \infty\right] \tag{7}$$

And the upper 1-sided 100×(1-α)% confidence interval,

$$\left[-\infty,\ \bar{Y} + t_{n-1,1-a} \times \frac{s_y}{\sqrt{n}}\right] \tag{8}$$

In the case of the alternative hypothesis in Eq. 5, a body of water will be listed only if k is *not* contained within the lower 1-sided 100×(1-α)% confidence interval. This can only occur if the lower bound is greater than k. Under these hypotheses, a mean that is considerably greater than k may still be judged to come from attaining waters if the lower bound of the confidence interval is less than k.

The upper 1-sided 100×(1-α)% confidence interval (Eq. 8) is appropriate for assessing attainment/impaired status whenever the hypothesis pairs in Eq. 6 are to be tested. In this scenario, only waters whose mean has an upper bound that is less than k will be classified as attaining the standard. Since, by definition all sample means that are greater than k must have upper 1-sided confidence limits > k, all bodies of water with sample means >k will be listed with at least 100×(1-α)% confidence. Thus regardless of whether one chooses to base their WQA decision on confidence intervals or on hypothesis tests, the null and alternative hypotheses of Eq.

6 will always lead to a decision wherein impaired waters will be listed with at least 100×(1-α)% confidence.

C.3.4. Data Quality Assessment Case History: Monitoring dissolved oxygen (DO) downstream from an agricultural operation

This example concerns the comparison of the proportion of samples that fail to meet a water quality criterion, to the maximum proportion allowed by state or federal water quality standards. The example is presented in a continuous format to show the five-step DQA process and its relationship to the DQO process that preceded it.

0. Background

In January 2000, a 3-year dissolved oxygen monitoring program was initiated at a sampling station located 0.60 miles downstream from a large commercial hog operation that had a history of manure lagoon overflows. This example demonstrates the application of the DQA process to make an interim decision based on a statistical analysis of data from the first full year of monitoring. In this case, the study was formulated, planned, and implemented through a rigorous DQO process that is illustrated in the DQO example on pages 15-22. Two additional DQA case histories will demonstrate how to apply DQA principles to the analysis of data from studies that were not designed through the DQO process.

1. Review the Study DQOs and Associated Sampling Design

Review the study Objectives. The basic research problem has been described in the preceding background discussion. The data for the first year included 34 DO measurements taken at 11-day intervals beginning on January 2, 2000. All of the scheduled evaluations were made; there were no missing data (Table 5). The unit of analysis is the area-adjusted mean DO measured at the study site on each of the 34 evaluations.

Translate the Objectives into Statistical Hypotheses: The baseline condition assumed that no more than 10% of the 34 samples had area-adjusted mean DO values < 5.0 mg/l. The alternative condition was that more than 10% of the samples failed to attain this DO criterion. Therefore the null and the alternative hypotheses are:

H_0: the proportion of the 34 samples with mean DO < 5.0 mg/l is ≤ 0.10
H_a: the proportion of the 34 samples with mean DO < 5.0 mg/l is > 0.10

The DQO process specified a gray region that was bounded on the left by the action level (0.10) and on the right by 0.25 (Fig. 4). A decision was made to accept decision errors within the gray region. For example, if one erroneously concluded that a set of samples with 15% of its area-adjusted mean DOs < 5.0 mg/l had attained the standard (10%), the resulting error would not have serious consequences because the proportion of attaining samples would still be quite high (85%). Outside the gray region, the acceptable false negative and false positive error rates were constrained to be ≤ 0.15 (Fig. 4). With a sample size of 34 and these specifications, the occurrence of 6 or more DOs < 5.0 mg/l (i.e., ≥17.6%) would result in rejection of the null hypothesis with a false positive rate of ≤ 0.12. Conversely, if 5 or fewer non-attainment means were observed, the null hypothesis would be accepted with a false negative rate of ≤ 0.15.

2. Conduct a preliminary Review of the Data

The 34 area-adjusted mean DO values are tabulated, by sampling date, in Table 5. DO values that are below the action level are marked with an asterisk. Twenty out of 34 means (59%) were below the action level. Except for a brief period in January – May, it appears that DO values at the study site regularly and frequently fell below the 5.0 mg/l action level.

A frequency distribution of the DO values is shown in Fig. 18a. The graph was made by subdividing the sample data into 14 groups based on the their DO values. Each bar is centered on the midpoint of the range of DO values in each group; the height of the bar denotes the number of samples in each subgroup. From inspection of Fig. 18a, it is obvious that the majority of the distribution of DO values is less than the 5.0 mg/l action level. There is little doubt that the10% non-attainment criterion was exceeded at the monitoring site during 2000. Nonetheless, it is desirable to incorporate some estimate of uncertainty (due to sampling error) into the decision-making process. A statistical test of hypotheses will be employed for this purpose.

3. Select the Statistical Test

Selecting the Test. Although the original data were means, the criterion is written in terms of the proportion of those means that do not attain the minimum DO concentration of the action level. The proportion itself is just the number of non-attainment means (20) divided by the total number of samples (34). This proportion (0.59) is compared against the criterion value (0.10). However, the comparison should take account of the uncertainty that arises from sampling the water at the station during 2000. In order to do this, a mathematical model of the random sampling variation in the proportion of non-attainments must be used. Because the proportion measures the occurrence of one of two possible outcomes for each sample (i.e., attainment or exceedance of the 0.10 criterion), the discrete binomial distribution is the appropriate model.

There are two statistical tests available to test the null hypothesis (attainment) vs. the alternative (exceedance): the exact binomial (binomial test) and the normal approximation (Z-test). Because the Z-test is based on an approximation to a discrete distribution by a continuous distribution, a continuity correction is sometimes made to the Z-test statistic before computing the p-value. The normal approximation gives very close agreement to the exact binomial for sample sizes greater than 50 and very poor agreement for sample sizes less than or equal to 20. Results for sample sizes between 20 and 50 are comparable but still differ by a significant amount. However, when the continuity correction is applied to Z-tests with n=21-50, the normal approximation and the exact binomial test results are virtually identical.

Power analyses conducted during the DQO indicated that a sample size of 26 would be sufficient when the uncorrected Z-test was used. However, when the exact binomial or the continuity corrected version of the Z-test was used, power analysis indicated a minimum sample size of 32. It was decided that 32 was the conservative choice; this number was increased to 34 to permit division of the year into 11-day sampling intervals. For illustrative purposes, both the exact

Table 5 2000 DISSOLVED OXYGEN MONITORING DATA

SAMPLE	DAY	DISSOLVED OXYGEN (MG/L)
1	02JAN	5.1
2	13JAN	4.5*
3	24JAN	10.3
4	04FEB	6.4
5	15FEB	8.2
6	26FEB	8.3
7	08MAR	7.3
8	19MAR	3.7*
9	30MAR	5.3
10	10APR	3.5*
11	21APR	5.1
12	02MAY	3.9*
13	13MAY	3.4*
14	24MAY	2.4*
15	04JUN	4.1*
16	15JUN	9.4
17	26JUN	6.2
18	07JUL	1.1*
19	18JUL	3.3*
20	29JUL	6.0
21	09AUG	5.0*
22	20AUG	3.2*
23	31AUG	8.5
24	11SEP	1.2*
25	22SEP	2.8*
26	03OCT	4.0*
27	14OCT	2.5*
28	25OCT	3.5*
29	05NOV	5.1
30	16NOV	7.2
31	27NOV	2.5*
32	08DEC	2.6*
33	19DEC	0.9*
34	30DEC	4.5*

binomial and Z-tests (with and without the continuity correction) well be used to test the null vs. the alternative hypothesis. Details of these tests are provided in Sections B2.4 and B2.5 of Appendix D.

Identify the Assumptions Underlying the Statistical Test. The model for the binomial distribution is appropriate for discrete response variables that meet to following assumptions:

1. The response can have only two outcomes (e.g., attainment, exceedance)
2. The underlying probability of exceedance, p, remains constant from sample to sample
3. Samples are obtained through an independent random sampling design

4. Verify the Assumptions of the Statistical Test

The exact binomial test is an example of a nonparametric test and as such does not require restrictive assumptions relating to the shape or the variability of the distribution. Thus no specific goodness of fit tests or graphical methods are needed to verify the assumptions. The three required assumptions can be verified by reviewing the sampling design and the data listing in Table 5. The response variable is clearly dichotomous. However, the second assumption is problematic. The binomial is most familiar as a model for the probability of obtaining various numbers of heads from repeated coin tosses. Assumption 2 essentially says that the binomial probability model requires that the very same coin be used for each toss. For example, if two different coins are used, one of which is fair (P=0.50) and other is not (p=0.75), the cumulative distribution function of the binomial will not provide correct estimates of the probabilities of obtaining specific numbers of heads. This assumption when applied to the DO sampling data requires that the exceedance probabilities at the monitoring site not change seasonally. Table 5 suggests that this assumption does not hold. Dissolved oxygen tends to be higher in colder months when less decomposition occurs. Thus the probability of exceedance is lower during the winter and early spring sampling periods than at other times of the year.

The assumption of independence would be valid if the assumption of constant-p were valid. However, because of the seasonality in exceedance probability, there is likely some autocorrelation in the data. In this case, the investigators were unable to make any specific adjustments or corrective actions to account for the violations of assumptions 2 and 3 , and simply proceeded to apply the exact binomial test to the data. In practice, when seasonality is observed in the data, a statistician experienced in the analysis of seasonal time series should be consulted to determine if an alternative approach or model should be employed.

5. Draw Conclusions from the Data

Perform the Statistical Hypothesis Test. The test statistic for the exact binomial distribution is just the number of samples out of 34 that had a DO value < 5.0 mg/l. As with all one-sample hypothesis tests we want to obtain a p-value that represents the probability of observing a test statistic greater than or equal to the one we have obtained from our data set. To do this, we need to know the expected distribution of the test statistic when the null hypothesis is true.

The probability distribution of the exact binomial for a population of 34 samples with an underlying 10% non-attainment probability is shown in Figure 18b. The histogram is the graphical representation of the discrete binomial probability density function. The total area within the darkened bars sums to 1.0. The probability that we desire is the blackened area that lies just to the right of the 20-sample point on the horizontal axis. This point is not shown in the figure because the there is no discernible probability of observing more than 9 exceedances out of 34. Table 6 lists the individual terms of the cumulative probability of the binomial distribution with n=34 and P=0.10. We can see that the probability of observing 13 or more exceedances is always <0.0001. In fact, the exact binomial probability of 20 or more non-attainment values from a population with only 10% non-attainment is 3.43×10^{-12}. Recall that we made a decision during the DQO process that if the probability of observing a value greater than or equal the sample test statistic was < 0.15 (i.e., we set α=0.15), we would reject the null hypothesis and conclude that the data did not support a decision that the water body attained the DO standard during 2000. Thus we must reject the null hypothesis that the true proportion of non-attainment at the monitoring site was ≤ 0.10.

The curve that delineates the normal approximation to the binomial with N=34 and p=0.10 is overlaid on the binomial probability distribution in Figure 18b. The total area under the bell-shaped curve is 1.0. The curve is smooth because it plots the probability distribution of a continuous, rather than a discrete, random variable. In this case, the random variable is Z:

$$z = \frac{Y - m}{{}^{s}\!/_{n}}$$

where $m = np$

y = the observed number of non-attainment values out of n samples

${}^{s}\!/_{n} = \sqrt{np(1-p)}$

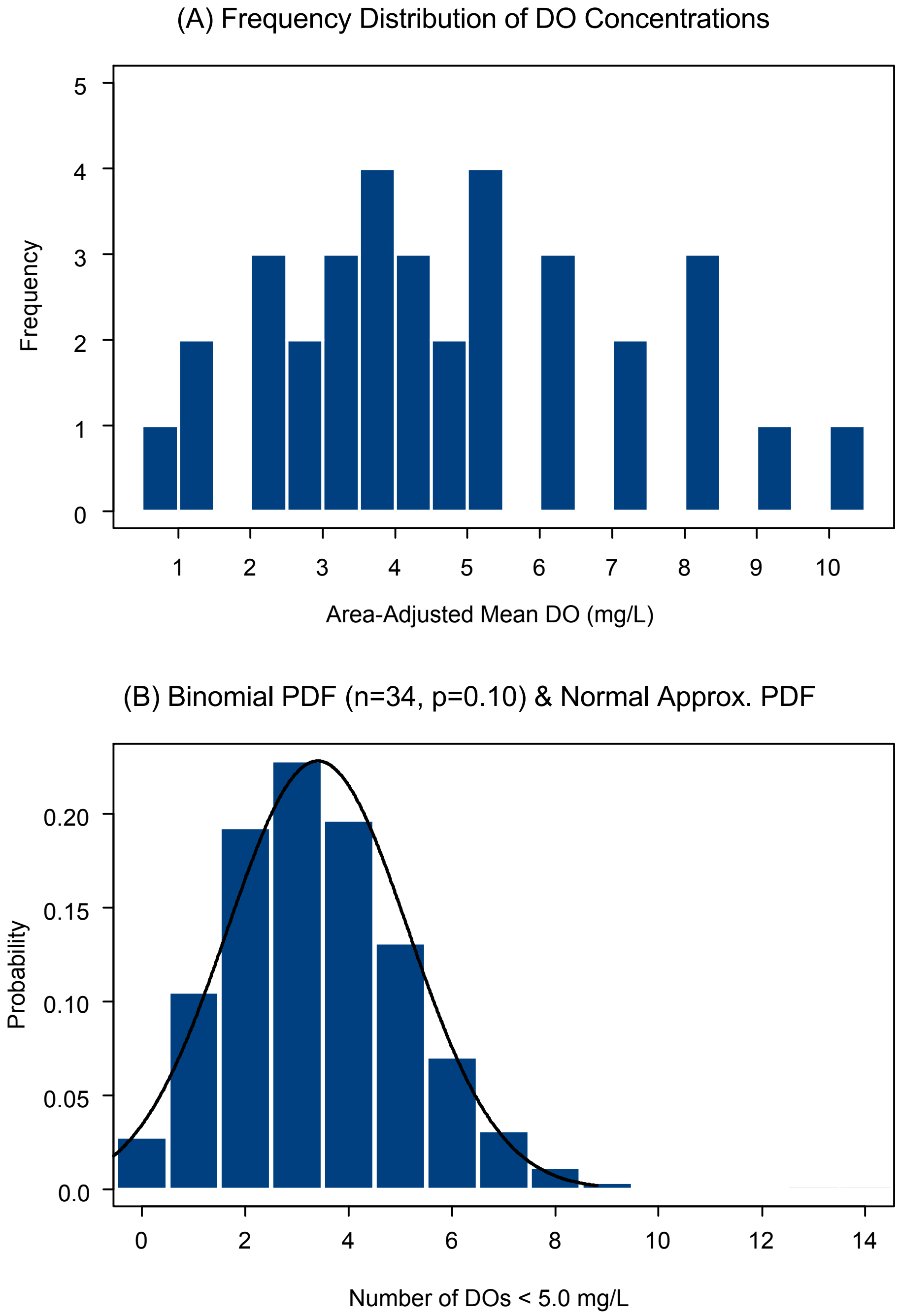
Substituting 34 for n and 0.10 for p, we get μ=3.4 and σ=1.75. Thus the Z for Y=20 non-attainment values is 9.49. This is an extremely large Z-score; 95% of the normal probability density lies between Z=± 2.0. Thus a value of Z as large as 9.49 has very little probability of occurring. In fact, its P-value is $< 10^{-25}$.

The continuity-corrected normal approximation for the upper-tailed, one-sided alternative hypothesis is calculated as:

$$z = \frac{Y - 0.50 - m}{{}^{s}\!/_{n}}$$

The result is Z=9.20 which likewise has $P < 10^{-25}$.

Fig. 18. (A) Frequency distribution of 34 area-adjusted DO means (mg/l) taken from the Mermantau River, at 11-day intervals, Jan-December 2000. (A) Relative frequency histogram and normal probability function (μ = 100.3, σ = 86.1). (B) Expected frequency distribution of n=34 means under $H_{0:}$ the distribution of exceedances is normal with μ = 3.4 and σ = 1.75.



```
TABLE 6 CUMULATIVE PROBABILITIES OF 1-34 SUCCESSES
             OUT OF N=34 TRIALS (P=0.10)

                   NUMBER       CUMULATIVE
                SAMPLES WITH     BINOMIAL
               DO < 5.0 MG/L    PROBABILITY

                      1            0.9722
                      2            0.8671
                      3            0.6745
                      4            0.4462
                      5            0.2496
                      6            0.1185
                      7            0.0481
                      8            0.0169
                      9            0.0051
                     10            0.0014
                     11            0.0003
                     12           <0.0001
                     13           <0.0001
                     14           <0.0001
                     15           <0.0001
                     16           <0.0001
                     17           <0.0001
                     18           <0.0001
                     19           <0.0001
                     20           <0.0001
                     21           <0.0001
                     22           <0.0001
                     23           <0.0001
                     24           <0.0001
                     25           <0.0001
                     26           <0.0001
                     27           <0.0001
                     28           <0.0001
                     29           <0.0001
                     30           <0.0001
                     31           <0.0001
                     32           <0.0001
                     33           <0.0001
                     34           <0.0001
```

The horizontal axis scale in Figure 18b is incremented by the number of observed non-attainments. This is the appropriate scale for the binomial distribution. The overlaid normal probability curve shows the probabilities associated with the z-transformation of the binomial scale values. For example, the value of 8 non-attainments on the binomial scale translates into (8-3.4)/1.75 = 2.63 on the Z-scale. The proportion of the normal curve to the right of 2.63 is the normal approximation of the area that lies to the right of 8 in the histogram of the exact binomial probability distribution. Similarly, the p-value of $<10^{-25}$ associated with the Z-score of 9.20 is an approximation of the exact binomial P-value of 3.43×10^{-12}. Both P-values lead to the rejection of H_0.

Draw Study Conclusions. Because the sample size (n=32) specified in the DQO insured that the false positive rate will be no greater than 0.15, we can be assured that the results, based on n=34, have an acceptably low probability of falsely rejecting the null hypothesis. In fact, the extremely large proportion of non-attainments (i.e., 0.59) places this estimate well beyond the upper boundary of the gray zone in Fig. 4 where the false negative rate is vanishingly small. Thus both the data (Table 5; Fig. 18 a) and the test statistics lead us to conclude that the dissolved oxygen content at the monitoring station was lower the 5.0 mg/l action level, more than 10% of the time during 2000. These results provide a basis for listing the reach of the river on which the hog operation is located and with the imposition of further corrective actions by the hog facility operators.

Evaluate Performance of the Sampling Design. Because this testing scenario involves a binomial proportion, the variability of the response was known exactly before the study was initiated. Thus it is not necessary to review the sampling design after the data were collected.

C.3.5. Data Quality Assessment Case History: Monitoring 3-Year Mean Turbidity in a River Reach

This example concerns the comparison of the sample mean value of a nonpriority water quality variable to its criterion value as set by state environmental regulations. The example is presented in a continuous format to show how to implement the five-step DQA process for a study that was not designed through the DQO process.

0. Background

Turbidity is the degree of opaqueness produced in water by suspended particulate matter. Turbidity may be quantified by measuring the degree to which light is scattered and/or absorbed by organic and inorganic material suspended in the water column. The standard unit of measure for turbidity is the nephelometric turbidity unit (NTU); larger NTU values indicate increased turbidity and decreased light penetration. The greater the amount of absorption and/or scattering, the less transparent the water, and the shallower the depth to which sunlight can penetrate. Because sunlight may only penetrate a few inches below the surface of waters with high turbidity, photosynthesis may become confined to the surface layers, leading to reductions in dissolved oxygen and productivity in deeper waters. Decreases in primary production associated with increases in sedimentation and turbidity may produce negative cascading effects through depleted food availability to zooplankton, insects, freshwater mollusks, and fish. Direct effects at each trophic level include mortality, reduced physiological function, and avoidance; however, decreases in available food at different trophic levels also result in depressed rates of growth, reproduction, and recruitment (Henley et al. 2000).

Water quality regulations in a fictitious parish in Louisiana required the 3-year mean turbidity of all lotic waters to be less than 150 NTU. Beginning in January 1980, the parish began to collect data on turbidity and other water quality parameters in a 20-mile reach of the Mermentau River. During this 20-year period, turbidity was recorded *in situ* using several different models of tubidimeters. The monthly turbidity data (sorted by NTU value) for 1980 - 2000 are shown in Table 3. The subset of the data for the 3-year period between 1997 and 1999 is shown in Table 1 of Appendix D. The sampling schedule was not designed through the application of a DQO process; rather it was based on a combination of budgetary considerations and past practices that were designed to capture seasonal variability.

1. Review the Study DQOs and Associated Sampling Design

Review the Study Objectives. The basic research problem was to estimate a 3-year mean turbidity value for a 20-mile reach of the Mermentau River and to determine whether, given the sampling error associated with the underlying sampling design, the 3-year mean was less than the state water quality standard of 150 NTU. The sampling design called for 21 turbidity measurements taken at 1-mile intervals along the river reach. Measurements were taken such that 10 readings were taken 10 meters from one shore and 11 were taken 10 meters from the opposite shore in a systematically alternating pattern. The near shoreline (i.e., left or right) chosen for the first measurement was alternated monthly. Operation and deployment of an YSI-6026 wiped turbidity sensor were made following the manufacturer's guidelines. Data were collected between 9:00 AM and 12:00 PM on the 15th day of each month or on the nearest Friday or Monday when the 15th fell on a Saturday or a Sunday. All scheduled collections were made; there are no missing data [Table 1 (Appendix D)]. The unit of analysis was the monthly mean turbidity (NTU) computed from each month's systematic sample of 21 turbidity measurements.

Translate the Objectives into Statistical Hypotheses: Because the parish regulations were written in terms of 3-year mean turbidity values, the population parameter of interest was the mean turbidity value of the specific 20-mile reach of the Mermentau River. Parish water quality standards required this mean to be less than 150 NTU. Therefore the hypotheses of interest are, "3-year mean turbidity ≥ 150 NTU" and "3-year mean turbidity < 150 NTU".

There are two possible decision errors: 1) to decide that the turbidity in the river reach exceeds the criterion value when in fact it does not; or 2) to decide that the turbidity in the river reach is less than the criterion value, when in fact it exceeds it. Because of the profound ecological and economic risks associated with sustained high turbidity (i.e., fish kills, undesirable algal blooms, proliferation of undesirable and possibly pathogenic microorganisms, poor visibility for navigation, fouling of industrial intake pipes, etc.), the consequences associated with a decision error of the second type were deemed to be far more serious than those of the first type. Accordingly, the null hypothesis was that the turbidity in the river reach *exceeded* the criterion value ("3-year mean turbidity ≥ 150 NTU") and the alternative hypothesis was that the mean turbidity attained the state water quality standards ("3-year mean turbidity < 150 NTU"). (See

Section C.3.3 for more information on choosing the most appropriate null and alternative hypotheses.)

Develop Limits on the Decision Errors: The gray region associated with an alternative hypothesis of the form, "3-year mean turbidity < 150 NTU" is bounded on the right by the action level (150 NTU) on an the left by a value that is less than the action level. Values of the 3-year mean that fall within the gray region are subject to high probabilities of falsely declaring a population 3-year turbidity mean that is < 150 NTU, to be > 150 NTU; this is called a "false-negative" decision error. The consequences of this type of error include unnecessary expenditure of resources on remedial actions, unnecessary regulatory action, and unnecessary economic hardship on individuals and companies that would be affected by erroneously listing the river and restricting its uses. Therefore we would like the gray region to be as narrow as possible. However, reducing the width of the gray region will require larger sample sizes (and hence greater expense) if the false negative error rate is to be maintained at a reasonably small value. With these cost-benefit relationships in mind, the lower bound on the gray region initially was set at 120 NTU (20% below the action level). It was decided that this value should be subject to revision if, after setting the minimum acceptable decision error rates, it required a sample size that was larger than the 36 monthly means that were available to make the decision.

Because the potential economic and social costs of falsely rejecting the null hypothesis were high, it was decided that the two types of decision error rates should be set equal to one another. Moreover, because the maximum sample size was already fixed at 36 and there was a desire to keep the width of the gray region as small as possible, it was decided that moderate sized error rates were all that was feasible. Therefore the false acceptance and false rejection error rates were both set to 0.15. Using the USEPA DEFT software with these specifications, it was found that a minimum of 63 samples would be required. Consequently, the lower bound of the gray region was extended to 110 NTU with the result that the DEFT software computed a new minimum sample size of 31. Thus, the given sample size of 36 insured a false negative error rate of ≤ 0.15 and a false positive rate of 0.15 with a gray region extending from 110 NTU – 150 NTU. These relationships are illustrated for the log-scale turbidity data in Figure 19.

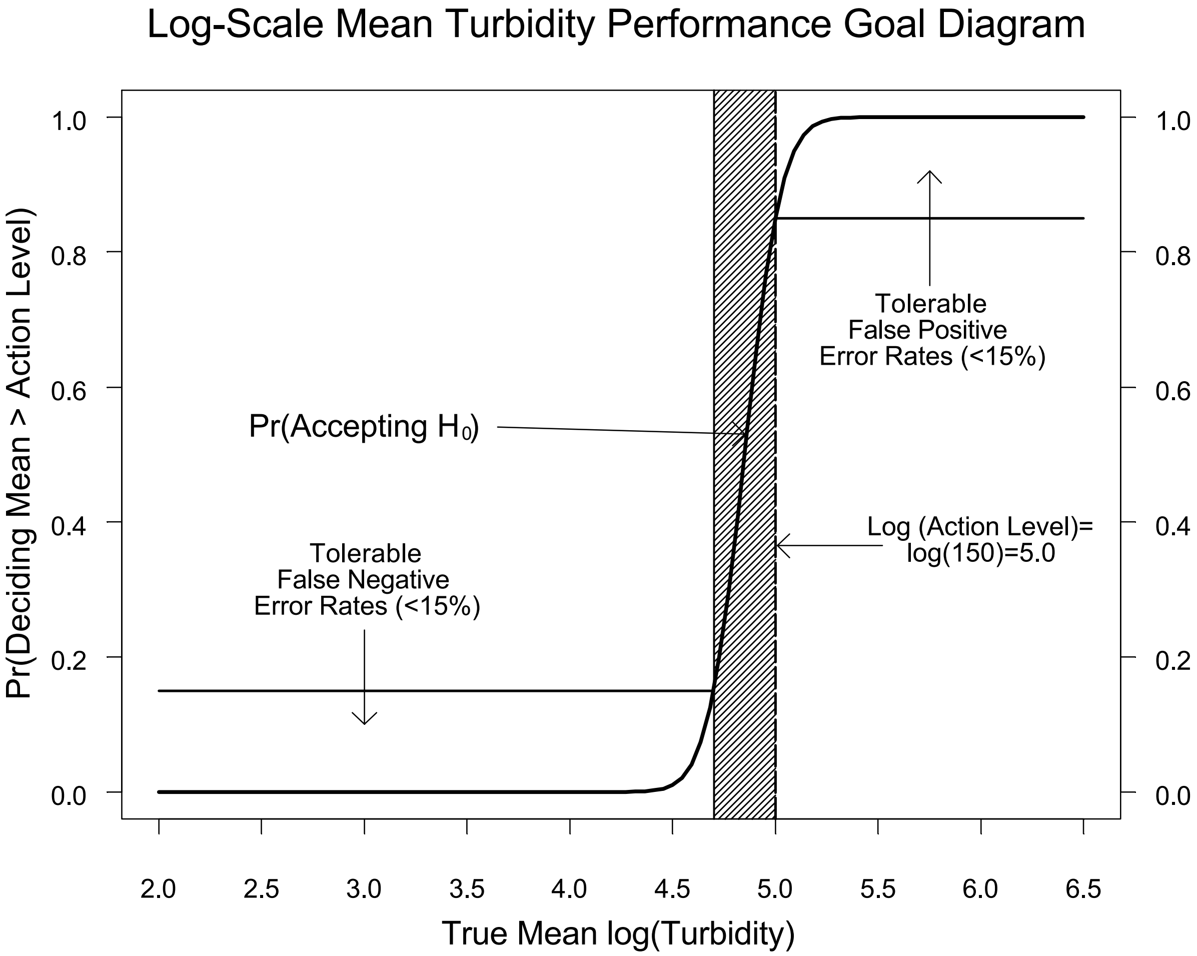
2. Conduct a preliminary Review of the Data

The 36 monthly mean turbidity values in NTU and in units of the natural log of the NTU values are shown in Table 1 (Appendix D). Previous analyses of these data revealed that they were lognormally distributed; thus, all of the calculations associated with the retrospective DQO steps have been based on consideration of the log-scale turbidity data. By inspection, there is clearly seasonality in the turbidity data and it appears that measurements taken in consecutive months are more alike than those taken farther apart in time. This suggests that some adjustment for seasonality and temporal autocorrelation may be required when the data are statistically analyzed.

3. Select the Statistical Test

Selecting the Test. Although the water quality criteria for turbidity were written in terms of a 3-year mean, we will make the assumption that the regulation was meant to apply to the center of

Fig. 19. Decision performance goal diagram for a one-sample t-test of H_0: population mean = 5.0 vs. H_a: population mean < 4.7, when the standard deviation =250, the sample size=30 and the width of the gray region is 0.30 (minimum detectable effect size=d=0.30 on the log-scale).



the distribution of the turbidity values in the target population. As explained in section C.2.2 (pages 38-40), the mean is interpretable as a measure of central tendency (i.e., as the median) when the data are normally or at least symmetrically distributed. Twenty years of monitoring in the Mermentau River and historical data from other nearby rivers had shown that monthly turbidity in these waters was consistently lognormally distributed. The back-transform of a mean computed on the log-scale is called the geometric mean. The geometric mean of lognormal data is an unbiased estimator of the population median on the original scale (Section C.2.2). Thus a statistical test that compares the log-scale mean turbidity to the natural log of the action value is equivalent to comparing the median on the original scale to the action level (i.e., 150 NTU). The following pair of hypotheses is appropriate for such a comparison:

$$H_0\text{: mean of log(Turbidity)} \geq \log(150)$$
$$H_a\text{: mean of log(Turbidity)} < \log(150)$$

The appropriate statistical test for these hypotheses is the one-sample t-test against a lower one-sided alternative (see Box 8 and Section D.2). These were the hypotheses of interest for the turbidity assessment and therefore the one-sided t-test was used to support the decision of attainment/nonattainment of the 150 NTU criterion by the 1998-1999 Mermentau turbidity data.

The presumption behind this approach is that the regulation for a maximum mean turbidity of 150 NTU was based on the equivalence of the mean and the median in normal distributions. However, if the intention of the regulators was in fact to focus on the mean regardless of whether it was located in the center of the distribution, then different estimation and testing procedures would be required. Chen's test can be used to test the above hypotheses on the means of the untransformed turbidity data and specialized estimation techniques are available for forming confidence intervals around the untransformed mean (Millard and Neerchal 2001).

Identify the Assumptions Underlying the Statistical Test. The one-sample t-test for lognormal data is appropriate for continuous response variables that meet to following assumptions:

a. The distribution of the natural logarithms of the data values are approximately normal
b. The data values come from an independent random sample of the target population

4. Verify the Assumptions of the Statistical Test

Assessment of the distributional form of the turbidity data was described in detail in Section C.2.2. Q-Q plots were used to verify that the log-transformed turbidity was approximately normally distributed (Fig. 6). In Section C.1.7, the independence assumption was shown not to hold for the 1997-1999 turbidity data; significant seasonality and temporal autocorrelation were demonstrated. However, a modification to the usual t-test calculation was applied to provide a means of testing the hypotheses of interest. The modification employed an adjusted estimate of the sample variance and a formula for adjusted degrees of freedom for the t-test. The adjusted variance takes account of the estimated autocorrelation in the monthly time series data. Because autocorrelated measurements tend to be alike, the 36 monthly turbidity measurements contain redundant information. This is reflected in the adjusted degrees of freedom that were computed for the t-test: df=8 (Appendix D, Section D.2). Because the degrees of freedom for the one-sample t-test = n-1, the effective sample size is only n=9 or 25% of the actual sample size. This creates a situation in which the specified maximum allowable false negative error rate of 0.15 is likely to be exceeded; i.e., the presence of significant autocorrelation decreases the effective samples size and the power of the statistical test.

5. Draw Conclusions from the Data

Perform the Statistical Hypothesis Test. The test statistic for the adjusted t-test is:

$$t_{a,df} = \frac{\overline{y} - k}{\frac{\sqrt{s^2_{adj}}}{\sqrt{nm}} \times \sqrt{\frac{1+\hat{f}_1}{1-\hat{f}}}}$$

where, $\overline{y}$ = the mean log-turbidity

k=log(150 NTU)

$t_{a,df}$ = the adjusted t-statistic with df degrees of freedom

n = the number of years of data

m = the number of months in a year

df = ((n×m)-m)/3

s^2_{adj} = the variance of the seasonally adjusted log-turbidity

$\hat{f}_1$ = the estimated autocorrelation in the seasonally adjusted log-turbidity at lag1

There were 12 monthly means in each of the three years of data, so df=8. The denominator of the above expression is the standard error of the 36-month turbidity time series. This expression is explained in Section D.2 of Appendix D; its value for the 1997-1999 data was 0.164 and the log-scale mean was 4.12, yielding an adjusted t-statistic of 5.43 with a P-value of 0.0003 for the lower one-sided alternative hypothesis that the true log-scale mean turbidity is < log(150) NTU. Because we have specified a false positive error rate of α=0.15 and the observed P-value is < α, we should reject the null hypothesis and accept the alternative hypothesis that the population 3-year median turbidity is < the 150 NTU action level. The actual false negative error rate associated with the effective sample size of 9 is β=<0.0001, substantially smaller than the planned rate of 0.15 that was associated with n=31.

Draw Study Conclusions. Because of the autocorrelation in the monthly turbidity measures, the effective sample size was only n=9, far below the minimum of n=31 that was specified in the DQO. This result demonstrates that even when the DQO steps are followed, if the independence assumption is violated, the estimated minimum sample sizes may be too small to maintain the desired decision error rates. In this case the geometric mean turbidity of the sample (61.6 NTU) was so far below the 150 NTU action value that even a sample size of n=9 had more than sufficient power to distinguish it from the action value. However, if the sample geometric mean had been closer to the action value, the false negative error rate (β) could easily have exceeded the value of 0.15 that was specified in the DQO. It is important to understand that the DQO estimates are contingent on the assumption that the conditions under which the DQO was carried out will prevail when the actual data are collected from the field. In this case that was not so, but as it turned out (rather fortuitously) the direction of the effects of these unforeseen conditions were favorable to maintaining decision error rates below the limits that were specified in the DQO.

Evaluate Performance of the Sampling Design. The results show that because of the substantial autocorrelation in the data, the sampling design employed was not adequate to provide data for a one-sided t-test with a specified false negative error rate ≤0.15. Unfortunately, restriction of the design to a 3-year period places severe limitations on what can be done. The decline of the effective sample size from 36 to 9 is a consequence of the seasonality and the autocorrelation in the turbidity time series. The usual remedy for inadequate sample size is to increase the number of samples that are collected, but in the case of the 3-year time series, this can only be done by increasing the sampling frequency within the 3-year period. This requires us to collect samples closer together in time, which will inevitably increase the autocorrelation. As the autocorrelation increases, the denominator term in the adjusted t-statistic shown above will get larger and the t-statistic will get smaller. Intuitively we can see that collecting more autocorrelated data, more frequently will not actually provide us with more useful information. Because of the redundancy in the data due to the autocorrelation, two measurements taken close together will simply be telling the same thing twice.

Ecologists and others who have analyzed annually fluctuating, seasonal environmental time series data have long recognized that decisions based on estimates and statistical tests of short-term environmental data (e.g., 3-year studies) may be prone to unacceptably high error rates. The only solution to the problem is to extend the duration of the study; some studies may extend for decades before it becomes possible to make good inferences from them. Generally speaking, once enough samples have been collected to estimate adequately the within-year seasonal effects (e.g., 12 monthly samples per year), the precision and decision error rates cannot be changed appreciably by simply increasing the number of samples within years. Thus, in cases where seasonality effects are pronounced and autocorrelation is high, it may not be possible to implement a DQO that will yield a sampling design that can ensure that specified decision error rates will be achieved within a 3-year period.

C.3.6. Data Quality Assessment Case History: Evaluating the 3-year Acute Maximum Concentration Criterion for Cadmium in a River Reach

This example concerns the comparison of the daily mean values of a priority toxic pollutant to its 3-year acute Criterion Maximum Concentration value (CMC) as set by EPA. The example is presented in a continuous format to show how to implement the five-step DQA process for a study that was not designed through the DQO process.

0. Background

Cadmium is a heavy metal whose chemical properties are similar to zinc. Cadmium does not occur uncombined in nature and is usually obtained as a byproduct of smelting and refining ores of zinc and lead. Cadmium is used principally for its anticorrosive properties in the electroplating of steel, in its sulfide form in the manufacture of paint pigments, and in the manufacture of batteries and other electrical components. Cadmium also occurs as a byproduct in many chemical fertilizers that are produced from phosphate ores. Cadmium enters the ambient air primarily from local smelting operations, it enters soil from local mining operations and from chemical fertilizers and it enters water from fertilizer runoff and/or industrial wastewater. Once in the water column, cadmium is rapidly absorbed by suspended particulates, which eventually settle into the sediments. From the sediments, it enters aquatic food chains where it tends to bioconcentrate in plants, crustaceans, and mollusks. When ingested by mammals, cadmium becomes concentrated in the liver and the kidneys.

Toxicity in mammals (including man) is due principally to kidney damage, which upsets calcium regulation and results in a net loss of calcium through the urine. Ultimately, this may lead to calcium depletion in bones and egg shells, which in turn may cause injury, reproductive failure, or death.

Mean concentrations of cadmium in unpolluted waters are typically < 1.0 μg/L. EPA has set the Criteria Maximum Concentration (CMC) of cadmium at 4.3 μg/L. The CMC, determined experimentally in laboratory studies, represents the highest concentration of a pollutant to which 95% of the freshwater aquatic organisms in an ecosystem can be exposed without suffering deleterious effects (Nowell and Resek 1994). If only one sample is taken in a 24-hour period, the pollutant concentration in that sample is the best estimate of the maximum concentration of the pollutant on that day, at that site. This value may be compared to the EPA CMC value to determine if the site attains the EPA standard on that day. However, EPA's CMC standards are written in terms of 3-year sampling periods. Specifically, the standard states that no more than one daily mean concentration may exceed the CMC for any given 3-year period at a single site. Theoretically, an aquatic ecosystem can recover from a single pollution event that exceeds the CMC during a 3-year period, but not from 2 or more such events. Thus, in practice, the total number of CMC exceedances for a specific pollutant (e.g., cadmium) in a 3-year period must be compared to the EPA standard of only one allowable CMC exceedance per 3-year monitoring period.

A fictitious parish in Louisiana routinely collects monthly samples of water from several lotic and lentic systems within the state. Concentrations of several pollutants are estimated from these samples and entered into an electronic database. Cadmium data collected during 1981-1983 from a reach of the Tangipahoa River in Southeastern Louisiana will be used in this case study to illustrate application of DQA procedures to the assessment of the once-in-3-years criterion for exceedances of the cadmium CMC. Because the monitoring program was designed prior to the development of EPA's DQO procedures, DQO principles will be applied retrospectively to frame the research question, formulate testable hypotheses, select an appropriate statistical test and establish bounds on the acceptable decision error rates.

1. Review the Study DQOs and Associated Sampling Design

Review the Study Objectives. The basic research problem was to determine whether, given the sampling error associated with the underlying sampling design, there was more than 1 daily exceedance of the cadmium CMC criterion (4.3 μg/L) in a reach of the Tangipahoa River, from January 1981- December 1983. Because the EPA CMC criterion for cadmium is written in terms of the number of days during a 3-year period that are allowed to have a daily mean cadmium concentration > the CMC, the target population is actually the 1095 days in the 3year period, each of which potentially could have had a daily mean cadmium concentration > 4.3 μg/L. Thus the primary objective of the study is to estimate the total number of days or, equivalently, the proportion of days during the three year period when the daily mean cadmium concentration was > 4.3 μg/L.

The sampling design called for one cadmium concentration measurement to be taken from a single sampling station in the 5-mile river reach within the first 3 days of each month during 1981-1983. A 200 ml sample of water collected from the monitoring station was filtered at the collection site. The filter with particulates and the filtrate were returned to the laboratory and subjected to further chemical analysis. In the laboratory, both particulates and filtrate were digested using 55% nitric and 70% perchloric acids in the ratio of 2:1. Cadmium concentrations were determined by means of flame atomic absorption spectrophotometry. Accuracy of each monthly measurement was insured by comparison to known analytical standards. The unit of analysis was the value of the mean cadmium concentration (μg/L) at the Tangipahoa monitoring station, on one day.

Translate the Objectives into Statistical Hypotheses: In order to answer the question of interest, "Was there more than 1 day during 1981-1983 when the mean daily cadmium concentration was greater than 4.3 μg/L?", we need an estimate of the total number of days with exceedant cadmium concentrations. The criterion for the cadmium CMC says that no more that 1 exceedance of the daily CMC can occur within a 3-year period. This leads naturally to the following pair of hypotheses:

H_0: total number of exceedant days ≤ 1
H_a: total number of exceedant days > 1

However, since the total number of days in a 3-year period is fixed (N=1095), the hypotheses could be framed in terms of the proportion of exceedant days:

H_0: proportion of exceedant days ≤ 1/1095 ≤ 0.0009
H_a: proportion of exceedant days > 1/1095 > 0.0009

The same probability model, the hypergeometric, can be used to evaluate both pairs of hypotheses (See Appendix D, Section D.6; Buonaccorsi 1987; Wendell and Schmee 2001). Like the binomial, the hypergeometric distribution model is appropriate for binary outcomes; e.g., attained vs., exceeded. The hypergeometric differs from the binomial model in that it assumes a finite population. For example, in the case of the CMC criterion, we are attempting to infer from a sample of 36 days to a "population" that contains exactly 1095 days. This contrasts with the use of the binomial distribution to infer from a small number of water samples to a river containing an essentially infinite number of such samples (e.g., see dissolved oxygen DQA case history).

For convenience, we will frame all statistical hypothesis tests and their associated error rates in terms of the proportion of exceedant days in a 3-year monitoring period. However, the hypergeometric distribution could easily be employed to test the equivalent hypotheses about the total number of exceedant days that occurred during the 3-year period. Given the second pair of hypotheses (above), there are two possible decision errors: 1) to decide that the proportion of daily mean cadmium concentrations >4.3 μg/L, exceeds 1/1095 when in fact it does not; or 2) to decide that the proportion of daily mean cadmium concentrations > 4.3 μg/L is less than or equal to 1/1095, when in fact it is not. Although the ecological and public health consequences of sustained high cadmium concentrations may be severe, the local economic and social consequences of incorrectly declaring the river reach to be in violation of the CMC once –in-3-years standard were considered more deserving of protection. Accordingly, the null hypothesis was that the proportion of exceedant days during the 1981-1983 monitoring period was ≤ 1/1095 and the alternative hypothesis was that the proportion of exceedant days was > 1/1095 (i.e., H_a: the mean daily cadmium concentration exceeded the CMC standard of 4.3 μg/L on more than 1 day during 1981-1983). (See Section C.3.3 for more information on choosing the most appropriate null and alternative hypotheses.)

Develop Limits on the Decision Errors: The gray region associated with an alternative hypothesis of the form, "The proportion of days during 1981-1983 with a mean cadmium concentration > 4.3 µg/L is > 1/1095" is bounded on the left by the action level (1/1095) on the right by a value that is greater than the action level (Fig. 20). Values of the proportion of exceedant days that fall within the gray region are subject to high probabilities of falsely declaring that the 3-year exceedance was < 1/1095 when in fact it was > 1/1095; this is called a "false-negative" decision error. The consequences of this type of error include possibly severe ecologic and pubic health effects. On the other hand EPA feels strongly that 2 or more exceedances in a year (i.e., ≥ 2/1095) should be cause for TMDL listing of the affected water body. Therefore it was decided that the gray region should be bounded on the left by 1/1095 and on the right by a proportion slightly less than 2/1095. Because 2/1095 = 0.00182, the upper bound was set at 0.0018. This means that the proportion associated with two exceedances will lie just outside of the gray and thus not be subject to excessively high false negative decision error rates.

Because the potential economic and social costs of falsely rejecting the null hypothesis were high, it was decided to set the Type I error rate at 0.05. The ecologic and public health consequences were deemed to be less important than the economic consequences so a higher Type II error rate of 0.20 was specified. The USEPA DEFT software does not compute samples sizes and power associated with hypergeometric probabilities. However these can be computed with other commercially available software (e.g., PASS). Specifying a lower bound of 1/1095 (0.00091) and an upper bound of 0.0018 for the gray region, a Type I error rate of 0.05, a Type II error rate of 0.20, and population size of 1095 days, PASS indicated that a minimum of 980 days out of 1095 would need to be sampled in order to meet the specifications. The reason for this large sample size is the extremely narrow gray region (width=0.0009); however, given the EPA standards which require listing for the occurrence of 2 or more exceedant days out 1095, there is no other acceptable choice for the gray region.

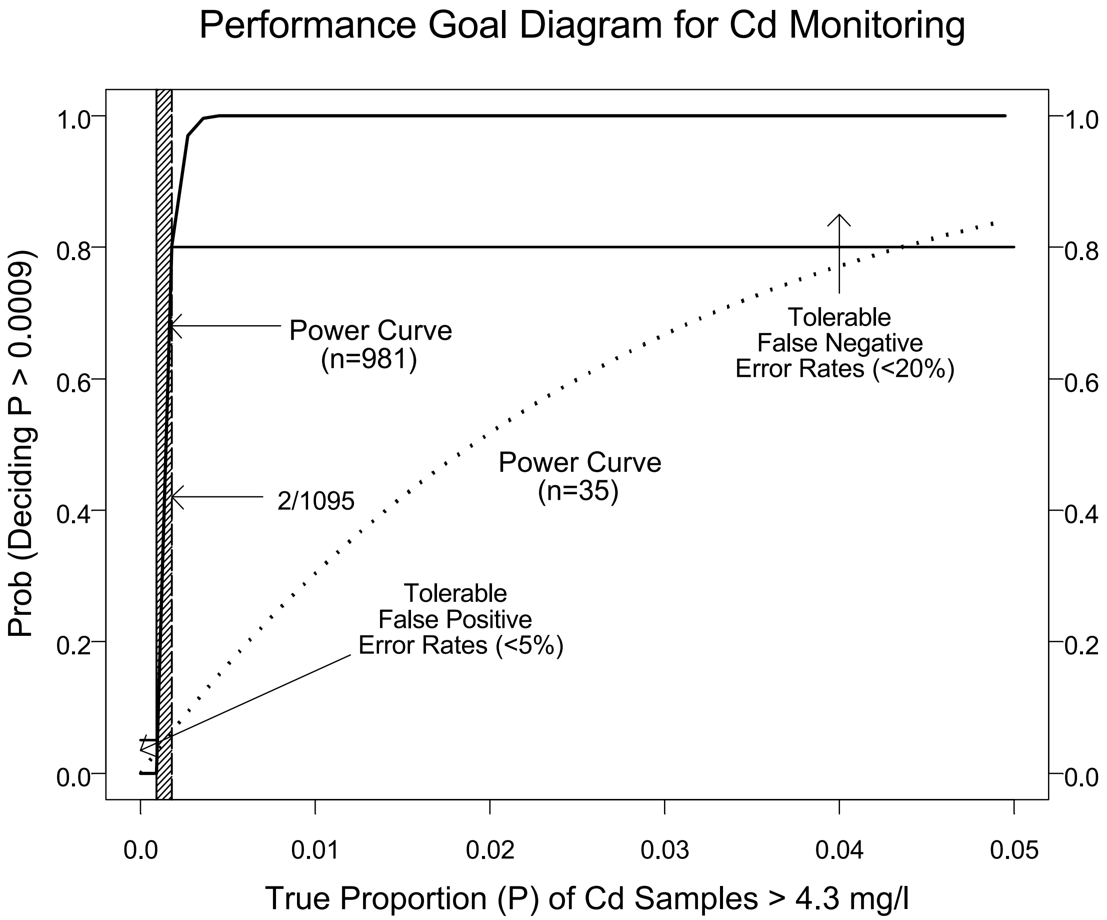
2. Conduct a preliminary Review of the Data

Only 35 of the scheduled 36 monthly sample measurements were available for analysis; the November 1982 sample was lost in transit to the laboratory. The 35 monthly cadmium concentration values are shown in Table 7. Although the two highest cadmium concentrations occurred in February, there is no clear pattern of seasonality or autocorrelation in the 1981-1983 cadmium data. Of the 35 concentrations, only one (5.0 µg/L in February 1981) exceeded the cadmium CMC. This suggests that the best estimate of the proportion of exceedant days during the 19981-1983 sampling period is the sample estimate, 1/35= 0.0286, a value that is clearly larger than the EPA standard of 0.0009 (1/1095). However, the point estimate does not account for the uncertainly due to the sampling. We will need to use the hypergeometric model in order to address the uncertainty in the sample estimate.

3. Select the Statistical Test

Selecting the Test. As noted above, the hypergeometric distribution is the appropriate probability model for binary outcomes from a finite population (i.e., a population of known fixed size, N). Although a one-sample z-test based on the normal approximation to the hypergeometric distribution is available (Cochran 1977), it does not perform well for sample sizes < 400 or when the population proportion is small (e.g., 1/1095; Johnson and Kotz 1969). Thus a decision was made to use the exact hypergeometric test to provide support for H_0 vs. H_a.

Fig. 20. Decision performance goal diagram for an exact hypergeometric test of H_0: the population 3-year exceedance rate is = 1/1095 vs. H_a: population mean > 1/1095, when the sample size=980 daily samples (solid power curve) or the sample size=35 daily samples (dotted line), and the width of the gray region is 1/1095 (minimum detectable effect size=d=1 daily exceedance).



If the upper tailed exact probability of observing a proportion larger than the sample proportion, is less than α (i.e., 0.05), the null hypothesis will be rejected and the population will be considered to have had more than one exceedant day during the monitoring period. Otherwise, the null hypothesis that the rate of exceedance is = 0.0009 (i.e., that there was not more than one exceedance in the river reach between Jan 1, 1981 and December 31, 1983) will be accepted.

Identify the Assumptions Underlying the Statistical Test. The Exact hypergeometric test is appropriate for a binary response (e.g., attains vs. exceeds) expressed as a proportion (e.g., the proportion of exceedances) or as a total count (e.g., total number of exceedances). The assumptions required for the exact hypergeometric test are:

1. The response can have only two outcomes (e.g., attainment, exceedance)
2. The underlying probability of exceedance, p, remains constant from sample to sample
3. Samples are obtained through an independent random sampling design.

4. Verify the Assumptions of the Statistical Test

The exact hypergeometric test is an example of a nonparametric test and as such does not require restrictive assumptions relating to the shape or the variability of the distribution. Thus no specific goodness of fit tests or graphical methods are needed to verify the assumptions. The three required assumptions can be verified by reviewing the sampling design and the data listing in Table 1. The response variable is clearly dichotomous. The second assumption would appear to be true for these data, at least there is no evidence in the Table 7 to suggest that the probability of exceedance deviates very much from zero over the course of the 3-year sampling period. Because the proportion of exceedances appears to be constant and near zero at all times, there is no suggestion that measurements taken close together in time are more likely to be exceedant than measurements taken at longer intervals; thus, the independence assumption appears to hold for these data.

5. Draw Conclusions from the Data

Perform the Statistical Hypothesis Test. The upper tailed hypergeometric probability was computed as 1-the cumulative hypergeometric probability function evaluated with r=1 (p=1/35), N=1095, n=35, and k=1 (P_0=1/1095). These values were substituted into Equation 4 of Appendix D to yield a cumulative probability of 1.0; thus the upper tailed p-value was zero. Since a=0.05 was specified, and the p-value was less than 0.05, the sample data did not support the null hypothesis that there was = 1 exceedant day in the river reach during 1981-1983. In fact, the corresponding estimate of the total number of exceedant days that occurred during 1981-1983 was 32 with a lower 1–sided 95% confidence limit of 2. Thus given the sampling error associated with the sample of n=35 days, we have 95% confidence that there were at least 2 days during 1981-1983 when the daily mean cadmium concentration exceeded the CMC. However, the false negative probability associated with this test is 0.937; i.e., in a 3-year period that actually had exactly 2 exceedant days, the probability that the sample estimate of the cumulative hypergeometric probability based on a sample of 35 days would lead to erroneous acceptance of the null hypothesis is 0.937. In plain English, this result says that if 1 exceedant day is found in a random sample of 35 days, there is a 95% probability that there was at least 1 additional day

TABLE 7 THE 1981-1984 CADMIUM DATA FROM THER TANGIPAHOA RIVER, LA

		CADMIUM
YEAR	MONTH	ug/L
1981	1	0.9
	2	5.0*
	3	3.7
	4	1.3
	5	2.7
	6	2.0
	7	2.7
	8	2.7
	9	2.4
	10	3.2
	11	2.1
	12	1.4
1982	1	2.0
	2	1.1
	3	1.8
	4	2.8
	5	1.1
	6	2.1
	7	1.6
	8	0.5
	9	1.2
	10	1.9
	12	0.6
1983	1	0.1
	2	3.9
	3	0.8
	4	0.2
	5	0.8
	6	1.1
	7	1.5
	8	0.8
	9	0.6
	10	0.2
	11	0.5
	12	0.4

* Concentration Exceeds the 4.3 mg/l Standard

that was exceedant during the 3-year period. Moreover, the high false negative rate tells us that unless the number of exceedant days (out of 1095) is much greater than 2, the exact hypergeometric probability estimate is not likely to indicate non-attainment. The null hypothesis was rejected in this case only because one exceedance in 35 days implies that there was likely a large number of exceedant days between January 1 1981 and December 31, 1983.

Draw Study Conclusions. The results of the exact hypergeometric test indicate a zero probability that the baseline assumption that there was ≤ 1 exceedant days in three years was true. This result indicates that that a serious problem with cadmium pollution likely occurred during 1983-1981. However, the very low power (0.063) associated with the test for a sample size of 35 is cause for concern for future studies in which the 1-in-3-years-exceedence criterion is to be based on similar sized samples.

Evaluate Performance of the Sampling Design. The low power associated with the once-per-month sampling design suggests that much larger sample sizes are required to protect against falsely concluding that there were less than 2 CMC exceedances in a 3-year period. In fact a power analysis based on the exact hypergeometric distributions shows that if we desire =0.05 and a false negative rate of 0.20, when there are as few as 2 exceedant days out of 1095 the minimum required sample size is 980 days; i.e., we would need to sample nearly every day to have an 80% chance of distinguishing whether there is more than one exceedance in a 3-year period. These relationships are illustrated in Fig. 20, which shows the associated power curves for the actual sample size (n=35 days) and the minimum sample size required to meet the DQOs (n=980). When n=980, and two exceedances actually occur within the 3-year period, the exact hypergeometric test has 80% power. However, when the sample size is 35 days a power of 80% is not achieved until 4.41% of the 1095 days (i.e., 49 days) are exceedant. Thus although EPA regulations require the listing of any water body with 2 or more exceedances, the sample size used for this study is reliable only for detecting exceedances that occur at a rate of 49 or more per 3 years of monitoring.

C.4. References

Browne, R.H. 1995. On the use of a pilot sample for sample size determination. Statistics in Medicine. 14:1933-1940.

Cleveland, W.S. 1993. Visualizing Data. Hobart Press. Summit, New Jersey, USA.

Elliot, J.M. 1977. Some methods for the statistical analysis of samples of benthic invertebrates. Freshwater Biological Assoc. Ambliside, Cumbria UK

Hengeveld, R. 1979. The analysis of spatial patterns of some ground beetles (Carabidae). Pages 333-346 *in:* M. Cormack and J.K. Ord, editors, Spatial and temporal analysis in ecology. International Co-operative Publishing House, Fairfield, Maryland, USA.

Henley, W.F., M.A. Patterson, R.J. Neves and A.D. Lemly. 2000. Effects of sedimentation and turbidity on lotic food webs: A concise review for natural resource managers. Reviews in Fisheries Science. 8(2):125-139

Millard, S. P. and N. K. Neerchal. 2000. Environmental Statistics with S-PLUS. CRC Press, Boca Raton, FL.

Nowell, L. and E.A. Resek. 1994. Summary of National Standards and Guidelines for Pesticides in Water, Bed Sediment, and Aquatic Organisms and their Application to Water-Quality Assessments. UGSS (http://ca.water.usgs.gov/pnsp/guide/guide_6.html)

O'Brien, R.G. 1998. A tour of UnifyPow: a SAS module/macro for sample size analysis. Proceedings of the 23rd SUGI Conference, 1346-1355. SAS Institute, Cary, NC.

Ord, J.K. 1979. Time series and spatial patterns in ecology. Pages 1-94 *in:* M. Cormack and J.K. Ord, editors, Spatial and temporal analysis in ecology. International Co-operative Publishing House, Fairfield, Maryland, USA.

Peterson, S.A., N.S. Urquhart, and E.B. Welch. 1999. Sample representativeness: a must for reliable regional lake condition estimates. Environ. Sci. Technol. 33:1559-1565.

Steel, G.D., J.H. Torrie and D.A. Dickey. 1996. Principles and procedures of statistics: a biometrical approach. McGraw-Hill, New York.

Stephan, C.E., D.I. Mount, D.J. Hansen, J.H. Gentile, G.A. Chapman, and W.A. Brungs. 1985. Guidelines for Deriving Numerical National Water Quality Criteria for the Protection of Aquatic Organisms and their Uses: EPA.

Thompson, S.K. 1992. Sampling. J. Wiley and Sons. New York, New York.

USEPA. 1999. National recommended water quality criteria – Correction. EPA 822-Z-99-001.

C.5 Glossary

alternative hypothesis - In a statistical hypothesis test there are two competing hypothesis, one of which, the alternative hypothesis, describes the set of conditions complementary to those described under the null hypothesis. For example: if the null hypothesis states that the mean pH of a set of samples is less than or equal to 5.0, the alternative hypothesis must be that the mean pH is greater than 5.0.

ANOVA Model - an acronym for Analysis of Variance. ANOVA models are linear models in which the total variance in a response is partitioned into two components: one due to treatments (and possible interactions among them) and the other due to random variability. In the simplest case where there is only one treatment factor, if the treatments have no effect on the response, the ratio of the variance components should be close to 1.0. If the treatments effect the response means, the ratio of the treatment component to the random component will be greater than one. Under the null hypothesis that the treatments have no effect, the sampling distribution of the ratio of the two variance components, each divided by their respective **degrees of freedom**, will be an F-distribution.

ARIMA Model - an acronym for autoregressive integrated moving average model. ARIMA models are linear models for regression and/or discrete treatment effects, measured through time, on responses that have been differenced at the appropriate **lag** distances.

autocorrelation - the internal correlation of a set of measurements taken over time and/or space. The correlation arises from the fact that points closer together in space and/or time tend to be more alike than those that are further apart. The autocorrelation function (either spatial or temporal) is a mathematical expression that relates the strength of the correlation to the distance (called the **lag**) between measurements.

Bayesian statistical inference - An approach to **inference** or **estimation** in which a process (e.g., a random binomial process) is proposed for the generation of a set of data. A mathematical model called a **likelihood** is specified for the process, such that the model parameters are random variables. A distribution, called the prior distribution, is developed for the parameters based on what is known about them, prior to collection of the data. Data are then collected and a mathematical principle called Bayes theorem is used to derive a second distribution of the parameters, called the posterior distribution, from the data and the prior distribution. The appropriate inference is then obtained from the posterior distribution. The Bayesian approach differs from the the classical frequentist approach in that is utilizes the investigator's prior knowledge of the system through the prior distribution.

bias - the systematic or persistent distortion of a measurement process that causes errors in one direction.

binary characteristic - a characteristic that can only have two possible values.

census - a study that involves the observation and/or measurement of every member of a population.

confidence interval - a range of values, calculated from the sample observations, that is believed, with a particular probability, to contain the true population parameter value. For example, a 95% confidence interval implies that if the estimation process were repeated again and again, then 95% of the calculated intervals would be expected to contain the true parameter value.

confidence level (also called the the confidence coefficient) - the probability that the confidence interval will include the true parameter value; Equivalently, 1-the probability (α)that the true value is *not* contained within the interval .

continuous random variable - A random variable which may take on an infinite number of values.

convenience sample - a sample collected from a target population without implementation of a probability-based design. Sampling units are selected based on ease of data collection, without clear reference to an underlying frame; e.g., the collection of water samples near bridges rather than randomly throughout the stream reach to which an inference is desired. Because many (perhaps the majority) of the population sampling units have little or no probability of selection to the sample and because sample coverage typically is restricted to some potentially homogeneous subset of the target population, data from convenience samples are not valid for statistical inference to the target population.

correlation coefficient – A scale-invariant measure of the association between 2 variables that takes on values between –1.0 and +1.0. The correlation coefficient has a value of plus one whenever an increase in the value of one variable is accompanied by an increase in the other, zero when there is no relationship (i.e., the 2 variables are independent of one another), and minus one (-1) when there is an exact inverse relationship between them.

correlogram - a plot or graph of the sample values of the autocorrelation coefficient of a time series against different values of the **lag**.

decision error - an error that occurs when data misleads an investigator into choosing the wrong response action, in the sense that a different action would have been taken if the investigator had access to unlimited "perfect data" or absolute truth. In a statistical test, decision errors are labeled as false rejection (Type I) or false acceptance (Type II) of a null hypothesis.

degrees of freedom (df) - As used in statistics, df has several interpretations. A sample of *n* values is said to have *n* degrees of freedom, but if *k* functions of the sample values are held constant, the number of degrees of freedom is reduced by *k*. In this case, the number of degrees of freedom is conceptually the number of independent observations in the sample, given that *k* functions are held constant. By extension, the distribution of a statistic based on *n* independent observations is said to have *n-p* degrees of freedom, where p is the number of parameters of the distribution.

discrete random variable - A random variable which may take on only a finite number of values.

dispersion - the amount by which a set of observations are spread out from their mean and/or median.

effect size - In a **one-sample test**, the difference between the sample mean and a pre-specified criterion or standard value. In a **two-sample test**, the effect size is the expected difference between the mean of a treatment group or ambient site vs. the mean of a control group or reference site. Associated statistical tests typically evaluate the null hypothesis of a zero effect size vs. the alternative that the effect size is nonzero.

effective sample size - When data are collected from cluster-correlated populations, there is redundancy in the information carried by more highly correlated individuals. Thus, correlated individuals carry less information than do uncorrelated individuals. The effective sample size is the number of uncorrelated individuals from a simple random sample that would carry information equivalent to the information in the sample of correlated individuals. The effective sample size is always less than the apparent sample size; how much less, is a function of the strength of the correlation and the sampling design that was used to collect the data.

estimation - the process of providing a numerical value for a population parameter on the basis of information collected from a sample.

experimental design - the arrangement or set of instructions used to randomize subjects to specific treatment or control groups in an experimental study. Such a procedure generally insures that results are not confounded with other factors and thus provides scientifically defensible inferences regarding causal effects of the treatments.

exploratory data analysis (EDA) - an approach to data analysis that may reveal structure and/or relationships among measured or observed variables in a data set. EDA emphasizes informal graphical procedures that typically are not based on prior assumptions about the structure of the data or on formal models.

extreme values - the largest and smallest values (and perhaps their neighboring values) among a sample of observations.

frequentist statistical inference - an approach to statistics based on the likelihood of an observed result in a large or infinite number of independent repetitions of the same sampling or experimental procedure (e.g., see the frequentist definition of the **confidence interval** in this glossary).

geometric mean - a measure of central tendency calculated by back-transforming the mean of a set of log-transformed observations. If the original data come from a log-normal distribution, the sample geometric mean will provide an unbiased estimate of the sample **median**.

heterogeneous - a term denoting inequality or dissimilarity of some quantity of interest (most often a variance) in a number of different groups, populations, etc.

homogeneous - a term denoting equality or similarity of some quantity of interest (most often a variance) in a number of different groups, populations, etc.

imprecision/precision - A term describing the degree of spread among successive estimates of a population parameter by a sample statistic. The standard error of a sample estimator (e.g., the standard error of the mean) is a measure of imprecision/precision in the estimator. A high degree of spread (imprecision) will lead to an increased likelihood of a decision error, while a reduction in spread will lead to a corresponding reduction in the likelihood of a decision error. Generally, precision will be increased by increasing the sample size.

independence - essentially, two events are said to be independent if knowing the outcome of one tells us nothing about the outcome of the other. More formally, two events *A* and *B* are said to be independent if Probability (A and B) = Prob(A) × Prob(B).

inference - the process of drawing conclusions about a population on the basis of measurements or observations made on a sample of individuals from the population.

lag - the distance, in units of time or space, between two events or locations. For example, an event occurring at time $t+k$ *(k>0)* is said to lag behind the event occurring at time *t*, by an amount of time equal to lag *k*.

likelihood - the probability of a set of observed data, given the value of some parameter or set of parameters associated with a model for the underlying process that is hypnotized to have generated the data. For example, if we obtain 9 heads in 10 tosses of a coin, the likelihood of observing this result, given that the coin is fair (i.e., the binomial parameter p=0.50), is approximately 0.0098.

log-transformation - a transformation on a variable, X, obtained as, Y=ln(X) or Y=ln(x+c), where c is a constant positive value (e.g., 1.0). This transformation is useful for normalizing continuous variables with skewed distributions and/or stabilizing the variance of a variable whose standard deviation increases as a function of its mean, prior to statistical analysis with tests (e.g., t-tests) that assume normality and/or variance homogeneity..

maximum likelihood - a procedure for estimating the value of a parameter(s) of a model of the underlying process that produced some particular set of observations, such that the resulting estimate maximizes the likelihood of the observed data. For example, the maximum likelihood estimate for the binomial parameter P, given an experiment in which one obtains 9 heads in 10 tosses is P=0.90. The likelihood of obtaining 9 heads given an underlying binomial process with P=0.90, is 0.3874. Note that the estimate P=0.90 leads to a much larger likelihood than an estimate of P=0.50 does (0.0098; see definition of **likelihood**). In fact there is no value of P that will yield a larger likelihood of obtaining 9 heads out of 10 tosses than the estimate P=0.90; thus, P=0.90 is the maximum likelihood estimator of P.

median - in a sample or a population, the median is the value of a random variable such that half of the sampling units have larger values and half have smaller values. When the population or sample size is 2N+1, the median is the value of the random variable associated with the N+1th ordered sampling unit; when the population or sample size is 2N, the median is average of random variable values of the sampling units with ranks N and N+1. If the population is normal the median will equal the mean. If the population is log-normal the median will equal the **geometric mean**.

Monte Carlo methods - methods for finding solutions to mathematical and statistical problems by simulation. Often used when the analytic solution of the problem is intractable, or when real data are difficult to obtain, or to evaluate the behavior of statistics or models under a variety of hypothetical conditions which may or may not be directly observable in nature.

nonparametric statistical methods (also called distribution-free methods) - Statistical techniques of **estimation** and **inference** are often based on the assumption of some underlying parametric process; for example, one that generates responses that are normally distributed. By contrast, nonparametric estimation and testing procedures do not depend on the existence of an underlying parametric process. Consequently, nonparametric techniques are valid under relatively general assumptions about the underlying population. Often such methods involve only the ranks of the observations rather than the observations themselves.

noncentral t-distribution - the expected distribution of the t statistic when the alternative hypothesis is true. This contrasts with central t-distribution (usually referred to simply as the "t-distribution") which is the expected distribution of the t-statistic when the null hypothesis is true. In general, the probability that an observed t-statistic comes from a non-central t-distribution will be large (e.g., P>0.20) when the probability of that it comes from a central t-distribution is low (e.g., P<0.001), and vice versa.

null hypothesis - a hypthesis about some presumed prevailing condition, usually associated with a statement of "no difference" or "no association" (see also **alternative hypothesis**).

one-sample tests - Statistical tests that evaluate the null hypothesis that there is no difference between a sample statistic (e.g., mean or proportion) and a fixed criterion or standard value.

parametric continuous distribution - the probability distribution of a continuous random variable, specified by a mathematical function of the population parameters; e.g., the normal distribution with parameters, μ and σ^2.

parametric statistical methods - tests and estimation procedures that depend on the complete specification of an underlying parametric probability distribution of the population from which the sample was drawn. The estimators and test statistics that are based on functions of the estimates of the population parameters under the assumed population distribution model (e.g. normal) are valid only if the assumed population model is valid. An example is the t-statistic which assumes an underlying normal population.

percentiles - the set of divisions of a set of data that produce exactly 100 equal parts in a series of values.

population - any finite or infinite collection of "units" that completely encompasses the set individuals of interest. In environmental studies, populations are usually bounded in space and time; e.g., the population of smallmouth bass in Leech Lake, Minnesota on July 1, 2000.

population parameter - a constant term(s) in a mathematic expression, such as a probability density function, that specifies the distribution of individual values in the population. Parameters typically control the location of the center of the distribution (location parameters), the spread of the distribution (scale or dispersion parameters) and various aspects of the shape (shape parameters) of the distribution (see also: **probability density function**).

power of a statistical test -the probability of rejecting the null hypothesis when it is false. Notice that we would like always to reject a false hypothesis; thus, statistical tests with high power (i.e., power >0.80) are desirable. Generally the power of a test increases with the number of individuals in the sample from which the test was computed.

precision - a term applied to the uncertainty in the estimate of a parameter. Measures of the precision of an estimate include its standard error and the confidence interval. Decreasing the value of either leads to increased precision of the estimator.

probability-based sample - a sample selected in such a manner that the probability of being included in the sample is known for every unit on the sampling frame. Strictly speaking, formal statistical inference is valid only for data that were collected in a probability sample.

probability density function (PDF)- for a continuous variable, a curve described by a mathematical formula which specifies, by way of areas under the curve, the probability that a variable falls within a particular range of values. For example, the normal probability density function of the continuous random variable X, is:

$$\frac{1}{s\sqrt{2p}}\exp\left[-\left(\frac{1}{2s^2}\right)(x-m)^2\right]$$

The normal probability density function has two **parameters**, the mean and variance , μ and σ^2. The mean is the location parameter and the variance is the scale parameter; the normal distribution does not have any shape parameters. The graph of the normal probability density function is the familiar “bell curve”.

rank - the relative position of a sample value within a sample.

relative frequency - the frequency of occurrence of a given type of individual or member of a group, expressed as a proportion of the total number of individuals in the population or sample that contains the groups. For example, the relative frequencies of 14 bass, 6 bluegill, and 10 catfish in a sample of 30 fish are, respectively: 46.7%, 20.0% and 33.3%.

representative sample - A sample which captures the essence of the population from which it was drawn; one which is typical with respect to the characteristics of interest, regardless of the manner in which it was chosen. While representativeness in this sense cannot be completely assured, probability-based samples are more likely to be representative than are judgement or convenience samples. This is true because only in probability sampling will every population element have a known probability of selection.

sample - a set of units or elements selected from a larger population, typically to be used for making inferences regarding that population.

sampling design - a protocol for the collection of samples from a population, wherein the number, type, location (spatial or temporal) and manner of selection of the units to be measured is specified.

sampling distribution - the expected probability distribution of the values of a statistic that have been calculated from a large number of random samples. For example, the sampling distribution of the ratios of each of the means from 100 samples (each with n=30) to their respective variances will be a t-distribution with 29 degrees of freedom.

sampling error - the difference between a sample estimate and the true population parameter due to random variability in the composition of the sample vs. that of the target population.

sampling frame - the list from which a sample of units or elements is selected.
sampling unit - the members of a population that may be selected for sampling.

significance level (α) - the level of probability at which it is agreed that the null hypothesis will be rejected; α is also the probability of a Type I error.

skewness - a measure of the asymmetry in a distribution, relative to its mean. A right-skewed distribution is composed mostly of small values lying close to the mean but possesses a few values that are much larger than the mean. Conversely, a left-skewed distribution is composed mostly of values lying close to the mean but possesses a few values that are much smaller than the mean.

square-root transformation - a transformation on a variable, X, obtained as, Y= $\sqrt{X}$ or Y= $\sqrt{X + \frac{1}{2}}$. This transformation is useful for normalizing a discrete variable with a Poisson distribution and/or stabilizing the variance of a variable whose variance is proportional to its mean, prior to statistical analysis with tests (e.g., t-tests) that assume normality and/or variance homogeneity.

standard deviation - the square root of the **variance**.

standard error - the standard error of a sample statistic, θ, (say a sample mean or proportion) is the standard deviation of the values of that statistic computed from repeated sampling of the target population, using the same sampling design (e.g., stratified simple random sampling) and the same sample size, *n*. For example, the standard error of the mean is the sample **standard deviation**/n.

standard normal distribution - a normal distribution whose mean is 0 and whose variance is 1.

statistic - a quantity calculated from the values in a sample (e.g., the sample mean or sample variance)**.**

statistical distribution - a probability distribution used to describe a statistic, a set of observations or a population.

statistical test of hypotheses - a statistical procedure for determining if a sample provides sufficient evidence to reject one statement regarding the population of interest (the null hypothesis) in favor of an alternative statement (the **alternative hypothesis**).

target population - the set of all units or elements, bounded in space and time, about which a sample is intended to produce inferences.

two-sample tests - Statistical tests that evaluate the null hypothesis that there is no difference between a sample statistic (e.g., mean or proportion) in a treatment group or at an ambient monitoring site and the sample statistic in a control group or at a reference site.

Type I error (α) - the error that occurs when a decision maker rejects the null hypothesis when it is actually true. Also called the false rejection decision error, or false positive decision error.

Type II error (β) - the error that occurs when a decision maker accepts a null hypothesis when it is actually false. This is also called the false acceptance decision error, or false negative decision error. The power of a statistical test is 1-β.

variance (population) - the variance of a finite population of N values - x_1,x_2,....x_N - is simply the average of the squared difference between the individual observations and the population mean.

variance (sample) - the variance of n sample observations is simply the average of the squared differences between the individual observations and the sample mean, divided by (n-1).

variogram - a plot of the sample values of the variance of a spatially referenced variable vs. the corresponding lag distances