
Chapter 9 - Quality Assurance: Design, Precision and Management

Quality assurance (QA) is an integrated program for ensuring the reliability of monitoring and measurement data and includes quality control. Quality control (QC) refers to operational procedures for obtaining prescribed standards of performance in the monitoring and measurement process. Specific QC elements can be developed for most, if not all, project activities. All project activities, from sampling (data collection) and laboratory analysis to statistical analysis and reporting, are potential error sources (Peters 1988). Because error is cumulative and can significantly affect the results of a project, all possible efforts must be made to control it. Therefore, quality assurance is a continuous process that should be implemented throughout the entire development and operation of a program.

The purpose of an overall quality assurance project plan (QAPP), containing specific QC elements and activities, is to minimize - and when possible eliminate - the potential for error. Additionally, there are objective mechanisms for evaluating activities relative to pre-established measurement quality objectives and other project goals. The appropriateness of the investigator's methods and procedures and the quality of the data to be obtained must be ensured before the results can be accepted and used in decision making. QA is accomplished through:

- Program design.
- Investigator training.
- Standardized data gathering and processing procedures.
- Verification of data reproducibility.
- Instrument calibration and maintenance.

As outlined below, QA requirements apply to all activities in an ecological study. More detailed guidance and examples for QA activities should be obtained from USEPA (1994d, 1995, 1996c); more general guidance is outlined by USEPA (1993b).

9.1 Program Design

A central component of QA is overall study design, which includes formulation of questions and hypotheses, experimental design, and development of analysis approaches. The classical approach by which scientists plan research consists of the following steps:

- Statement of the problem to be resolved.
- Formulation of alternative hypotheses that will explain the phenomena or, in the case of problems that do not involve elaboration of processes, formulation of specific research questions.
- Establishment of boundaries within which to resolve the problem.

- Formulation of an experimental or study design that will falsify one or more hypotheses or answer the specific research questions.
- Establishment of uncertainty limits including setting acceptable probabilities of Type I and Type II errors for statistical hypothesis testing.
- Optimization of the study design including power analysis of the statistical design.

Experimental advances in basic sciences have not included the last two steps because uncertainty limits were inappropriate or unknown. Examination of experimental advances also reveals that a high degree of creativity and insight is required to formulate hypotheses and study designs; no formal planning process or "cookbook" can guarantee creativity and insight. Nevertheless, documentation of the planning process and a complete explanation of the conceptual framework help others evaluate the validity of scientific and technical achievements.

9.1.1 Specifying the Questions

The first task in developing a sampling and assessment program is to determine, and be able to state in simple fashion, the principal questions that the sampling program will answer. Questions may or may not be framed as hypotheses to test, depending on program objectives. For example, suppose that a sampling program objective is to establish reference conditions for biological criteria for lakes in state Y. Typically, the initial objectives of a survey designed to develop criteria are to identify and characterize classes of reference lakes. Initial questions may then include:

- Should state Y's minimally disturbed lakes be divided into two or more classes that differ in biological characteristics and dynamics?
- What are the physical, chemical, and relevant biotic characteristics of each of the lake classes?

After state Y's monitoring and assessment program has developed biological criteria, new questions need to be developed that encompass assessments of individual lakes, groups of lakes, or lakes of an entire region or state. Specific questions may include:

- Is lake Z similar to reference lakes of its class (unimpaired), or is it different from reference lakes (altered or impaired)?
- Overall, what is the status of lakes in state Y? How many (or what percentage) lakes are similar to reference conditions? How many lakes are impaired?
- Has lake Z changed over a certain period? Has it improved or deteriorated?
- Overall, have lakes in state Y improved or deteriorated over a certain period? Have individual lakes improved? Are more lakes similar to reference conditions now than some time ago?

Finally, resource managers often wish to determine the relationships among variables, that is, to develop predictive, empirical (statistical) models that can be used to design management responses to perceived problems. Examples of specific questions include:

- Can trophic state of a lake be predicted by areal phosphorus loading rate (e.g., Vollenweider 1968)?

- Can the biota of a lake be predicted by watershed land use (e.g., Dillon et al. 1994)?

These same models (e.g., analysis of variance, regression) are also used to help develop hypotheses on causal relationships between stressors and responses of systems. Establishing cause requires manipulative experiments, and since surveys and monitoring programs preclude experimental investigations, inference of causal relations will not be considered here. Often, there is enough experimental evidence available from other studies so that additional causal experiments are not necessary and would be superfluous (e.g., current knowledge of nutrients and trophic state generally makes it unnecessary to "prove" experimentally which nutrients are limiting).

9.1.2 Specifying the Population and Sample Unit

Sampling is statistically expressed as a sample from a population of objects. In some cases, the population is finite, countable, and easy to specify, e.g., all lakes in state Y, where each lake is a single member of the population. In other cases, the population is more difficult to specify and may be infinite, e.g., lake waters of state Y, where any location in any lake defines a potential member of the population (Thompson 1992). Sampling units may be natural units (entire lakes, cobbles in a littoral zone), or they may be arbitrary (plot, quadrat, sampling gear area or volume) (Pielou 1977). Finite populations may be sampled with corresponding natural sample units, but often the sample unit (like a lake) is too large to measure in its entirety, and it must be characterized with one or more second stage samples of the sampling gear (bottles, benthic grabs, quadrats, etc.)

In most sampling designs, each sample unit is assumed to be independent of other sample units. The objective of sampling is to best characterize individual sample units in order to estimate some attributes (e.g., number of taxa, DO) and the statistical parameters (e.g., mean, median, variance, percentiles) of a population of sample units. The objective of the analysis is to be able to say something (estimate) about the population. It is critical to distinguish between making an inference about a population of many lakes (e.g., "Reservoirs in the Blue Ridge are deep and oligotrophic") versus an inference about a single lake (e.g., "Lake Z has fewer fish species than unimpaired reference lakes"). These two kinds of inferences require different sampling designs: the first requires independent observations of many lakes and does not require repeated observations within sample units (pseu-doreplication) (Hurlbert 1984); while the second often does require repeated observations within a lake. Table 9-1 depicts some examples of sample units and populations.

Sample Unit	Sample Population	Infinite or Finite Population
A point in a specific lake.	All points in the lake	Infinite
A point in <i>any</i> lake of a state or region.	Total surface area or volume in a state or region	Infinite
A lake or a definable	All lakes in a state or	Finite

subbasin of a lake as a single unit. (NOTE: Because lakes are most often discrete environments, this is likely to be the most common sample unit)	region	
--	--------	--

9.1.3 Specifying the Reporting Unit

Finally, it is necessary to specify the units for which results will be reported. Usually, these units are the population (e.g., all lakes), but often subpopulations (e.g., lakes within a given lake district) and even individual locations (e.g., lakes of special interest) can be used. Subpopulations, or strata, are more homogeneous than the entire population, and are separated to facilitate comparison among them (see Section 9.2.1). In order to help develop the sampling plan, it is useful to create hypothetical statements of results in the way that they will be reported, for example:

- *Status of a place:* Lake Z is degraded.
- *Status of a region:* 20% of the lake area in state Y has an elevated trophic state, above reference expectations; or 20% of lakes in state Y have an elevated trophic state.
- *Trends at a place:* Benthic species richness in lake Z has decreased by 20% since 1980.
- *Trends of a region:* Average lake trophic state in state Y has increased by 20% since 1980; or Average benthic index values in 20% of lakes of state Y have increased by 15% or more since 1980.
- *Relationships among variables:* 50% increase of P loading above natural background is associated with decline in number of taxa of benthic macroinvertebrates, below reference expectations; or Lakes receiving runoff from large impervious parking lots have 50% greater probability of elevated trophic state above reference than lakes not receiving such runoff.

Specification of reporting units helps to focus the study design on relevant questions. Alternative designs can be examined for their ability to address the questions within the specified reporting units. Elements of the design that are not relevant to questions and reporting units are identified as superfluous.

9.2 Sampling Design

9.2.1 Sources of Variability

Variability of data justifies the existence of statistics. Variability has many possible sources. The intent of sampling designs is to collect a representative sample of the population. For bioassessment, we also wish to (1) minimize variability due to uncontrolled measurement error and, (2) characterize and partition the natural variability. For example, we may stratify lakes by soil phosphorus content of the surrounding

watersheds (e.g., Rohm et al. 1995) so that lakes within a soil P class may be likely to have similar water column total P concentrations. Typically, we stratify so that observations (sample units) from the same stratum will be more similar to each other than to sample units in other strata.

When sampling lakes we often measure something (say, chlorophyll concentrations) at single points in space and time (center of the lake, 2m depth, 10 AM on 2 July). If we make the same measurement at a different place (littoral zone, 1 m) or time (30 January), the measured value will be different. These two natural components of variability (space and time in this example) are called sample variability or sampling error (Fore et al. 1994). A third component of variability, called measurement error, refers to our ability to accurately measure the quantity we are interested in. Measurement error can be affected by sampling gear, instrumentation, errors in proper adherence to field and laboratory protocols, and the choice of methods used in making determinations. The three basic rules of efficient sampling and measurement are:

1. Sample so as to minimize measurement error.
2. Characterize the components of variability that have influence on the central questions and reporting units.
3. Control other sources of variability that are not of interest and thus minimize their effects in the observations.

In our example of chlorophyll concentrations, we may want to sample each of several lakes in the deepest part, with a vertically integrated pump sample taken in early spring before stratification appears. Many lakes are sampled in order to examine and characterize the variability due to different lakes (the sampling unit). Each lake is sampled in the same way, in the same place, and in the same time frame in an attempt to minimize variability due to location, depth, and season, which are not of interest in this particular study.

In the above example, chlorophyll concentrations vary with location within a lake, among lakes, and time of sampling (day, season, year). If the spatial and temporal components of variability within lakes are large, then it is best to use either an index period sample or to estimate a composite from several determinations. For example, measurements of chlorophyll concentrations typically vary more between spring and fall samples within a lake than they do between lakes. Therefore, lake chlorophyll concentrations are often estimated as a growing season average, taken from several determinations (for instance, monthly) during the growing season.

In analyses, especially hypothesis testing, multiple determinations within lakes may be a form of pseudoreplication (Hurlbert 1984), and should be used with caution. If the hypothesis refers to a single lake (e.g., chlorophyll concentration of lake Z is higher than a biocriterion), multiple determinations are often necessary for the test. If the hypothesis refers to many lakes (e.g., lakes in state Y have elevated chlorophyll compared to state Q), multiple determinations within lakes are pseudoreplication if they are used as independent observations in the test, rendering the test invalid (Hurlbert 1984). If multiple determinations for each lake are used to calculate a single seasonal mean or median, which is then used as an independent observation for the hypothesis test, there is no pseudoreplication. Repeated measurement designs - analysis of variance (ANOVA

- ANalysis Of VAriance) or regression - can be used (e.g., Underwood 1994) as a single analysis that takes into account multiple determinations. These methods estimate means of repeated measures to maintain independence.

A less costly alternative to multiple measures in space is to use spatially composite determinations. In nutrient or chlorophyll determinations, a water column pumped sample, where the pump hose is lowered through the water column, is an example of a spatially composite determination. Benthic macroinvertebrates are often sampled with spatial composite determinations. For example, benthic macroinvertebrates in Atlantic Coastal Plain streams are typically sampled by 20 sweeps of a dip net in multiple habitats, and composited into a single sample (e.g. USEPA 1997b, Barbour et al. 1996a, Barbour et al. 1996b, Roth et al. 1997). Benthic sampling of Florida lakes is a composite of 12 Petite Ponar grabs made throughout the sublittoral zone of a lake or a sample unit (Gerritsen and White 1997) (see Florida case study in this chapter).

Multiple observations within a sample unit (e.g., within a lake) should not be considered independent observations unless they are taken to examine an explanatory variable of interest, such as effects of depth, lake zone, season, or year. The principal use of multiple measurements is to estimate measurement error, that is, the variability we should expect when a single determination is made in a lake.

Analysis of variance is used to estimate measurement error. All multiple observations of a variable are used (from all lakes with multiple observations), and lakes are the primary effect variable. The root mean square error (RMSE) of the ANOVA is the estimated standard deviation of repeated observations within lakes. A hypothesis test (F-test) is not of interest in this application because it tests the trivial hypothesis that lakes are different from one another.

Measurement error is the result of methodological biases and errors: gear bias; improper use of gear or improper training; variability in use of gear; laboratory errors (chemical analysis errors); and natural variability that is not of interest and is not being sampled. Measurement error is minimized with methodological standardization: selection of cost-effective, low variability sampling methods; proper training of personnel; and quality assurance procedures designed to minimize methodological errors.

Natural variability that is not of interest for the questions being asked, but may affect ability to address these questions, should be estimated with the RMSE method above. If the variance estimated from RMSE is unacceptably large (i.e., as large or larger than variance expected among sample units), then it is often necessary to alter the sampling protocol, usually by increasing sampling effort in some way, to further reduce the measurement error. Measurement error can be reduced by multiple observations at each sample unit, e.g.: multiple Ponar casts at each sampling event, multiple observations in time during a growing season or index period, depth-integrated samples, or spatially integrated samples.

Spatial integration of sample material and compositing the material into a single sample is almost always more cost-effective than retaining separate, multiple observations. This is especially true for relatively costly laboratory analyses such as organic contaminants and benthic macroinvertebrates. The Florida invertebrate and TVA fish methodologies

include the compositing of multiple sampler casts into a single sample, which is then counted and identified.

For quality assurance, some effort will always be required for repeated samples so that measurement error can always be estimated from a subset of sites. Repeated measurement at 10% or more of sites is common among many monitoring programs, and is recommended.

9.2.2 Alternative Sampling Designs

Sampling design is the selection of a part of a population in order to observe the attributes of interest, so that the values of those attributes can be estimated for the whole population. Classical sampling design makes assumptions about the variables of interest; in particular, it assumes that the values are fixed (but unknown) for each member of the population, until that member is observed (Thompson 1992). This assumption is perfectly reasonable for some variables, say, length, weight, and sex of members of an animal population, but it seems less reasonable for more dynamic variables such as nutrient concentrations, loadings, or chlorophyll concentrations of lakes. Designs that assume that the observed variables are themselves random variables are model-based designs, where prior knowledge or assumptions (a model) are used to select sample units.

9.2.3 Probability-based Designs (Random Sampling)

The most basic probability-based design is simple random sampling, where all possible sample units in the population have the same probability of being selected, that is, all possible combinations of n sample units have the equal probability of selection from among the N units in the population. If the population N is finite and not excessively large, a list can be made of the N units, and a sample of n units is randomly selected from the list. This is termed list frame sampling. If the population is very large or infinite (such as locations in a lake), one can select a set of n random (x,y) coordinates for the sample.

All sample combinations are equally likely in simple random sampling. There is no assurance that the sample actually selected will be representative of the population. Other unbiased sampling designs that attempt to acquire a more representative sample include stratified, systematic, multistage, and adaptive designs. In stratified sampling, the population is subdivided or partitioned into strata, and each stratum is sampled separately. Typically, partitioning is done so as to make each stratum more homogeneous than the overall population. For example, lakes could be stratified by ecoregion. Systematic sampling is the methodical selection of every k th unit of the population from one or more randomly selected starting units, and ensures that samples are not clumped in one region of the sample space. Multistage sampling requires selection of a sample of primary units, such as fields or hydrologic units, and then selection of secondary sample units such as plots or lakes within each primary unit in the first stage sample.

Estimation of statistical parameters requires weighting of the data with inclusion probabilities (the probability that a given unit of the population will be in the sample)

specified by the sampling design. In simple random sampling, inclusion probabilities are by definition equal, and no corrections are necessary. Stratified sampling requires weighting by the inclusion probabilities of each stratum. Unbiased estimators have been developed for specific sampling designs, and can be found in sampling textbooks, such as Thompson (1992).

9.2.4 Model-based Designs

Use of probability-based sampling designs may miss relationships among variables (models), especially if there is a regression-type relationship between an explanatory and a response variable. As an example, elucidation of lake response to phosphorus loading with the Vollenweider model (e.g., Dillon and Rigler 1974) required a range of trophic states from ultraoligotrophic to hypereutrophic. A random sample of lakes is not likely to capture the entire range (i.e., there would be a large cluster of mesotrophic lakes with few at high or low ends of the trophic scale), and the random sample may be biased with respect to the regression model.

In model-based designs, sites are selected based on prior knowledge of auxiliary variables, such as estimated phosphorus loading, lake depth, elevation, etc. Model-based designs may preclude an unbiased estimate of the population (e.g., regional trophic state), unless the model can be demonstrated to be robust and predictive. The population value is then predicted from the model and from prior knowledge of the auxiliary (predictive) variables.

Identifying and sampling selected least stressed reference sites to develop an index is an example of samples for a model. The model is the index (e.g., IBI) and the responses of its component metrics. Reference sites alone cannot later be used for unbiased estimation of the biological status of lakes. Ideally, it may be possible to specify a design that allows both unbiased estimation of a population and index or model development. Statisticians should be consulted in developing the sample design for a biocriteria and biological monitoring program. However, managers should be aware that there is strong disagreement among statistical schools of thought on the subject of sampling design.

9.3 Evaluation of Statistical Power

A principal aspect of probability sampling is determining how many samples will be required to achieve the monitoring goals and what is the probability of making an incorrect decision based on the monitoring results. The primary tool for conducting these analyses is statistical power analysis. Evaluating statistical power is key to developing data quality criteria and performance specifications for decision making (USEPA 1996c) as well as evaluating the performance of existing monitoring programs (USEPA 1992d). Power analysis provides an evaluation of the ability to detect statistically significant differences in a measured monitoring variable. The importance of this analysis can be seen by examining the possible outcomes of a statistical test. The null hypothesis (H_0) is the root of hypothesis testing. Traditionally, null hypotheses are statements of no change, no effect, or no difference. For example, the mean abundance at a test site is equal to the mean abundance of the reference sites. The alternative hypothesis (H_a) is

counter to H_0 , traditionally being statements of change, effect, or difference. Upon rejecting H_0 , H_a would be accepted.

The two types of decision errors that could be made in hypothesis testing are depicted in Table 9-2. A Type I error (i.e., false positive) occurs when H_0 is rejected although H_0 is really true. A Type II error (i.e., false negative) occurs when H_0 is accepted although H_0 is really false. The magnitude of a Type I error is represented by α and the magnitude of a Type II error is represented by β . Decision errors are the result of measurement and sampling design errors that were described in Section 9.2.1. A proper balance between sampling and measurement errors should be maintained because accuracy limits effective sample size and vice versa (Blalock, 1979).

Table 9-2. Errors in hypothesis testing.

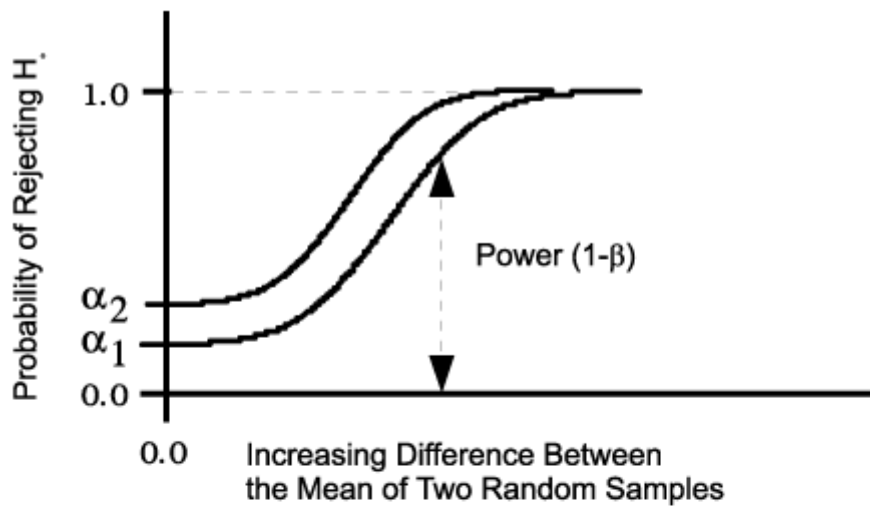
Decision	State of affairs in the population	
	H_0 is True	H_0 is False
Accept H_0	$1-\alpha$ (Confidence level)	β (Type II error)
Reject H_0	α (Significance level) (Type I error)	$1-\beta$ (Power)

9.3.1 Comparison of Significance Level and Power

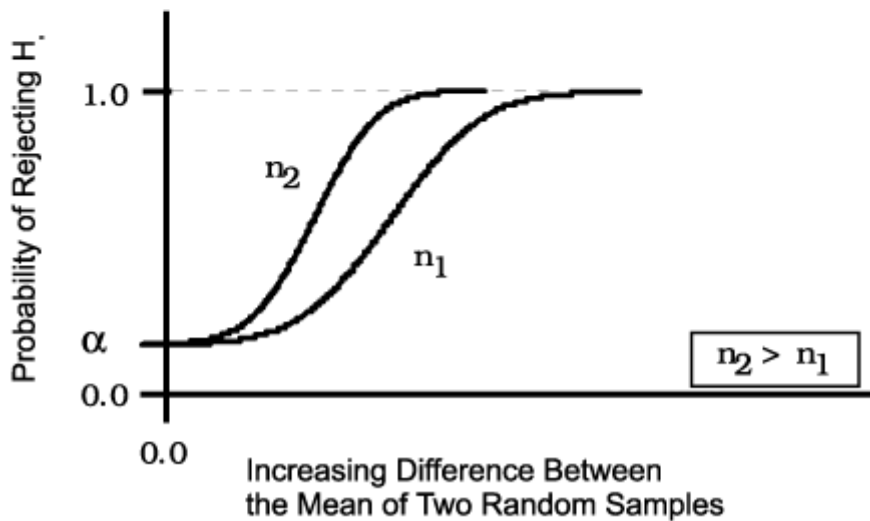
Regardless of the statistical test chosen for analyzing the data, the analyst must select the significance level of the test. That is, the analyst must determine what error level is acceptable. The probability of making a Type I error is equal to the significance level (α) of the test and is selected by the data analyst. In many cases, managers or analysts define $1-\alpha$ to be in the range of 0.90 to 0.99 (e.g., a confidence level of 90 to 99 percent), although there have been environmental applications where $1-\alpha$ has been set to 0.80. Selecting a 95 percent confidence level implies that the analyst will reject the H_0 when H_0 is really true (i.e., a false positive) 5 percent of the time.

Type II error depends on the significance level, sample size, number of replicates, variability, and which alternative hypothesis is true. The power of a test ($1-\beta$) is defined as the probability of correctly rejecting H_0 when H_0 is false. In general, for a fixed sample size, α and β vary inversely. Power can be increased (β can be reduced) by increasing the sample size or number of replicates. Figure 9-1 illustrates this relationship. Suppose the interest is in testing whether there is a significant difference between the means from two independent random samples. As the difference in the two sample means increases (as indicated on the x-axis), the probability of rejecting H_0 , the power, increases. If the real difference between the two sample means is zero, the probability of rejecting H_0 is equal to the significance level, α . Figure 1A shows the general relationship between α and β if α is changed. Figure 1B shows the relationship between α and β if the sample size is increased. The tradition of 95% confidence ($\alpha = 0.05$) is entirely arbitrary; there is no scientific requirement that confidence be set at

95%. Indeed, for environmental protection, power is at least as important - and possibly more important - than confidence (Peterman 1990, Fairweather 1991).



A) Increasing Significance Level from α_1 to α_2



B) Increasing Sample Size from n_1 to n_2

Figure 9-1. Illustration of significance (a) and power (1- β).

9.3.2 Basic Assumptions

Usually, several assumptions regarding data distribution and variability must be made to determine the sample size. Applying any of the equations described in this chapter is

difficult when no historical data set exists to quantify initial estimates of proportions, standard deviations, means, or coefficients of variation. To estimate these parameters, Cochran (1963) recommends four sources:

- Existing information on the same population or a similar population.
- A two-step sample. Use the first-step sampling results to estimate the needed factors, for best design, of the second step. Use data from both steps to estimate the final precision of the characteristic(s) sampled.
- A "pilot study" on a "convenient" or "meaningful" subsample. Use the results to estimate the needed factors. Here the results of the pilot study generally cannot be used in the calculation of the final precision because often the pilot sample is not representative of the entire population to be sampled.
- Informed judgment, or an educated guess.

For evaluating existing programs, proportions, standard deviations, means, etc. would be estimated from actual data.

Some assumptions might result in sample size estimates that are too high or too low. Depending on the sampling cost and cost for not sampling enough data, it must be decided whether to make conservative or "best-value" assumptions. Because of the fixed mobilization costs, it is probably cheaper to collect a few extra samples the first time than to realize later that additional data are needed. In most cases, the analyst should probably consider evaluating a range of assumptions regarding the impact of sample size and overall program cost. USEPA recommends that if the analyst lacks a background in statistics, he/she should consult with a trained statistician to be certain that the approach, design, and assumptions are appropriate to the task at hand.

9.3.3 Simple Comparison of Proportions and Means from Two Samples

The proportion (e.g., percent dominant taxon) or mean (e.g., mean number of EPT taxa) of two data sets can be compared with a number of statistical tests including the parametric two-sample t-test, the nonparametric Mann-Whitney test, and two-sample test for proportions (USEPA 1996c). In this case, two independent random samples are taken and a hypothesis test is used to determine whether there has been a significant change. To compute sample sizes for comparing two proportions, p_1 and p_2 , it is necessary to provide a best estimate for p_1 and p_2 , as well as specifying the significance level and power ($1-\beta$). Recall that power is equal to the probability of rejecting H_0 when H_0 is false. Given this information, the analyst substitutes these values into the following equation (Snedecor and Cochran, 1980)

$$n_b = (Z_a + Z_{2\beta})^2 \frac{(p_1 + p_2 q_2)}{(p_2 - p_1)^2}$$

Equation 1.

where Z and $Z_{2\beta}$ correspond to the normal deviate. Common values of $(Z + Z_{2\beta})^2$ are summarized in Table 9-3. To account for p_1 and p_2 being estimated, t could be substituted for Z . In lieu of an iterative calculation, Snedecor and Cochran (1980)

propose the following approach: (1) compute n_0 using Equation 1; (2) round n_0 up to the next highest integer, f ; and (3) multiply n_0 by $(f+3)/(f+1)$ to derive the final estimate of n .

Table 9-3. Common values of $(Z_\alpha + Z_{2\beta})^2$ for estimating sample size for use with Equations 1 and 2 (Snedecor and Cochran 1980).

Power, $1-\beta$	α for One-sided Test			α for Two-sided Test		
	0.01	0.05	0.10	0.01	0.05	0.10
0.80	10.04	6.18	4.51	11.68	7.85	6.18
0.85	11.31	7.19	5.37	13.05	8.98	7.19
0.90	13.02	8.56	6.57	14.88	10.51	8.56
0.95	15.77	10.82	8.56	17.81	12.99	10.82
0.99	21.65	15.77	13.02	24.03	18.37	15.77

To compare the mean from two random samples to detect a change of (i.e., $x_2 - x_1$), the following equation is used:

$$n_0 = (Z_\alpha + Z_{2\beta})^2 \frac{(s_1^2 + s_2^2)}{\delta^2}$$

Equation 2.

where s_1 and s_2 are standard deviation of samples 1 and 2.

Common values of $(Z + Z_{2\beta})^2$ are summarized in Table 9-3. To account for s_1 and s_2 being estimated, Z should be replaced with t . In lieu of an iterative calculation, Snedecor and Cochran (1980) propose the following approach: (1) compute n_0 using Equation 2; (2) round n_0 up to the next highest integer, f ; and (3) multiply n_0 by $(f+3)/(f+1)$ to derive the final estimate of n .

A special case of Equation 2 arises for biocriteria, when we compare the mean of a sample from a lake to determine if the value is below some set limit, that is, if the lake is impaired or below a reference threshold. The threshold is fixed by previous investigations and decisions, and is not a random variable. We ask now whether we can detect a change of (i.e., $C - x_1$), where C is the biocriteria limit:

$$n_0 = (Z_\alpha + Z_{2\beta})^2 \frac{(s_1^2)}{\delta^2}$$

Equation 3.

In Equation 3, Z is most often one-tailed, because the concern is only whether the value is below the threshold.

9.3.4 Sample Size Calculations for Means and Proportions

For large sample sizes or samples that are normally distributed, symmetric confidence intervals for the mean are appropriate. This is because the distribution of the sample mean will approach a normal distribution even if the data from which the mean is estimated are not normally distributed. The Student's t statistic ($t_{\alpha/2, n-1}$) is used to compute symmetric confidence intervals for the population mean, μ :

$$\bar{x} - t_{\alpha/2, n-1} \sqrt{s^2/n} \leq \mu \leq \bar{x} + t_{\alpha/2, n-1} \sqrt{s^2/n}$$

Equation 4.

where \bar{x} is the sample mean and s^2 is the sample variance.

This equation is appropriate if the samples are normally distributed or the sample size is greater than 30 (Wonnacott and Wonnacott, 1972), although Helsel and Hirsch (1992) suggest that highly skewed data might require more than 100 observations.

Although several approaches exist to estimate confidence levels for any percentile, many rely on assuming a normal or lognormal distribution. The approach presented here (Conover, 1980) for more than 20 observations does not rely on these assumptions. Conover (1980) also provides a procedure for smaller sample sizes.

Example Sample Size Calculations for Comparing Proportions and Population Means

Example 1 - Sample size calculation for comparing proportions

To detect a difference in proportions of 0.20 with a two-sided test, α equal to 0.05, $1-\beta$ equal to 0.90, and an estimate of p_1 and p_2 equal to 0.4 and 0.6, n_0 is computed from Equation 1 as

$$n_0 = 10.51 \frac{[(0.4)(0.6) + (0.6)(0.4)]}{(0.6-0.4)^2} = 126.1$$

Rounding 126.1 to the next highest integer, n is equal to 127, and n is computed as $126.1 \times 130/128$ or 128.1. Therefore 129 samples in each random sample, or 258 total samples, are needed to detect a difference in proportions of 0.2. Since these are proportions, the result means that the total count in the sample must be at least 129. For example, to detect the above difference in the proportion of dominant taxon (e.g., benthic macroinvertebrates or fish) of two lakes, at least 129 individuals must be counted and identified in each lake. The example illustrates that a statistically significant difference can be easily detected in proportions if sufficient individuals are sampled. However, it is doubtful that a difference between 40% and 60% in dominant taxon is biologically meaningful.

Example 2 - Sample size calculation for comparing population means

To detect a difference of 20 in mean abundance with a two-sided test. The standard deviation, s , was estimated as 30 for both samples based on previous studies; α was selected as 0.05; and $1-\beta$ was selected as 0.90. Substituting these values into Equation 2 yields

Rounding 47.3 to the next highest integer, f is equal to 48, and n is computed as $47.3 \times 51/49$ or 49.2. Therefore 50 samples in each random sample, or 100 total samples, are needed to detect a difference of 20.

To calculate the confidence interval corresponding to the median, lower quartile, or upper quartile, the following procedure is used.

1. Order the data from smallest to largest observation such that where x_p corresponds to the median (i.e., $p=0.5$), lower quartile (i.e., $p=0.25$), or upper quartile (i.e., $p=0.75$).

$$x_1 \leq \dots \leq x_r \leq \dots \leq x_p \leq \dots \leq x_s \leq \dots \leq x_n$$

2. Compute the values of r^* and s^* as Equation 5. where $Z_{\alpha/2}$ is selected from a normal distribution table.

$$r^* = np - Z_{\alpha/2} (np(1-p))^{0.5}$$

$$s^* = np + Z_{\alpha/2} (np(1-p))^{0.5}$$

Equation 5.

3. Round r^* and s^* up to the next highest integers r and s . The 1- lower and upper confidence limits for x_p are x_r and x_s , respectively.

It can be seen from Equation 5 that estimation of medians or quartiles from small samples can result in large confidence intervals for the estimate. For example, the 90% confidence interval for the lower quartile of a sample of $n=10$ covers the first five observations. A sample of less than 10 observations would have a confidence interval extending below the smallest observation. This is the reasoning behind a general "rule of thumb" that estimation of reference conditions should be based on a sample of 10 or more sites, if at all possible.

Case Study: Optimization of Benthic Sampling in Florida Lakes

To optimize its lake sampling protocols, Florida DEP performed a pilot study on 9 lakes. Each lake was sampled with twelve petite Ponar grabs (0.02 m²) distributed approximately equidistant in the sublittoral zone of the lake (2-4 m depth; Fig. 7-6). Each grab was kept separate in laboratory identification and enumeration. The lakes spanned a wide range in benthic

macroinvertebrate diversity and abundance (Table 9-4), from 7 to 63 taxa, and 228 to 3540 organisms in 0.24 m² sampled. In seven of the nine lakes, the number of taxa continued to increase with sampling effort, and did not reach an asymptote with twelve Ponar samples.

Table 9-4. Number of taxa and individuals in 12 cumulative Ponar samples from 9 Florida lakes.

Lake	Cumulative taxa	Cumulative individuals
Overstreet	63	768
Post	54	454
Camel	54	3540
Logan	42	1649
Mic	34	2828
Oche	31	1849
Del	16	228
Pickett	9	370
Adams	7	495

To illustrate the effects of compositing sample casts, each sample of 12 grabs was composited into 2 replicate samples of 6 casts, so that each sample consisted of alternate casts (Fig. 7-6). This yielded 2 alternative sampling protocols: 12 Ponar replicates for each lake, and two replicates of 6 Ponars each. 4 candidate metrics were calculated: number of taxa (cumulative for composited samples), percent dominance, sensitive taxa (ephemeroptera, trichoptera, odonata), and log abundance. Standard deviation of each metric, as measurement error in determining the "true" value for each lake, was estimated with the root mean square error (RMSE) from an analysis of variance (Table 9-5).

Table 9-5. Comparison of two sample processing protocols, Florida lakes.

	mean of 12 Ponars				mean of 2 samples of 6 composited Ponars			
	Population mean (9 lakes)	Range (9 lakes)	s.d. (individual lake)	CV (average lake)	Population mean (9 lakes)	Range (9 lakes)	s.d. (individual lake)	CV (average lake)
No. of taxa	8.85	2-19	3.62	40.9%	25.7	5.5-44.5	4.36	16.9%
% dominance	58.8%	40%-96%	14.8%	25.2%	50.4%	16%-96%	8.9%	17.7%
Sensitive taxa (ETO)	0.39	0-1.7	0.628	161%	1.6	0-5.5	1.27	79.4%
Total indiv (ln)	4.13	2.78-5.60	0.717	17.4%	6.12	4.68-7.48	0.145	2.4%

All metrics had a lower coefficient of variation (CV) in the composited protocol than in the uncomposited, showing the advantages of compositing multiple deployments of small sample gear such as Ponars. Composited samples reduce costs because fewer jars and records are required, and sampling time is reduced some. Laboratory analysis can be reduced by subsampling a fixed number of organisms (e.g., 100, 200, or 300) from the composite sample for identification. It has been shown with the same Florida data (Barbour and Gerritsen 1996) that subsampling a fixed number of organisms (100 or more) yields adequate estimates of number of taxa, which are actually more precise than taxa density (total taxa in a fixed area or volume) (Hurlbert 1971). Subsamples that are larger than the target number can be reduced computationally by rarefaction (Hurlbert 1971, Vinson and Hawkins 1996, Barbour and Gerritsen 1996). Based on these results, Florida DEP adopted the following sampling protocol for lake benthic invertebrates:

- 12 Ponars randomly deployed in 12 segments of the 2-4m depth zone of lakes less than 1000 acres.
- Ponar casts are composited into a single sample and sieved through a 500 m mesh screen.
- A subsample of 100 benthic macroinvertebrates is sorted and identified to the lowest practical taxonomic level.
- The sampling protocol is duplicated at approximately 10% of sites to estimate measurement error.

Case Study: Estimation of Power for TVA Fish Samples

TVA samples reservoir fish, benthic macro- invertebrates, water column chlorophyll, dissolved oxygen, and sediment contamination to rate the overall health of its reservoirs. 5 indices are calculated, one for each indicator group. Measurements are duplicated at selected reservoirs to obtain estimates of variability.

In 1996, fish sampling was repeated at seven reservoirs. The TVA Reservoir Fish Assembly Index (RFAI) is composed of 12 metrics (see

Chapter 8). Ranges of metric values in 1996 (for all reservoirs) and metric standard deviations (from multiple determinations at single reservoirs are given in Table 9-6.

From the standard deviation of the RFAI score, we can estimate the number of samples required to detect differences among lakes.

1. Difference between two lakes (or between two sampling times within a lake)

To detect a difference of 10 in mean RFAI score with a two-sided test. The standard deviation, s , was estimated as 4.027 for both samples (Table 9-6); α was selected as 0.05; and $1-\beta$ was selected as 0.80. Substituting these values into Equation 2 yields

Rounding 2.54 to the next highest integer, f is equal to 3, and n is computed as $2.54 \times 6/4$ or 3.82. Therefore, 4 samples in each reservoir, or 8 total samples, are needed to detect a difference of 10 in RFAI score between two reservoirs, with a probability of 0.80 of finding a true difference.

2. Test whether a lake is below a threshold (biocriteria)

To detect a difference of 10 in mean RFAI score below a threshold. The same standard deviation estimate is used as above (4.027; Table 9-6); and $1-\beta$ were selected as 0.05 and 0.80, respectively, but is now one-sided. Substituting these values into Equation 2 yields:

Table 9-6. Minimum and maximum values, and standard deviations of repeated measures, of reservoir fish metrics and the RFAI.

Metric	Minimum (all reservoirs, 1990-96)	Maximum (all)	s of repeated measures (n=7)
Total taxa	12	47	1.389
Piscivore species			1.309
Sunfish species			0.756
Sucker species			0.463
Intolerant species			0.463
Percent tolerant			0.118
Percent dominant			0.122
Percent omnivores			0.118
Percent insectivores			0.141
Simple lithophil species			0.463
Total individuals			15.677
Percent anomalies			0.0051
RFAI index score	18	58	4.027

Rounding 1.002 to the next highest integer, f is equal to 2, and n is computed as $1.002 \times 5/3$ or 1.67. Therefore, 2 samples are needed to detect a difference of 10 in RFAI score below a threshold, with a probability of 0.80 of finding a true difference. If the effect size, or distance below the

threshold, were increased to 15, then the required sample size would be 1. Thus if we find an RFAI value from a single unreplicated sample to be 15 points below a threshold, then we would expect that replication would not change a conclusion that the reservoir RFAI is below the threshold, 95% of the time. This example shows the potential value of adaptive sampling strategies, where a decision to increase sampling effort is based on the value of the first replicate. If the index value is very far below a threshold, there is no need to replicate. As the index value approaches the threshold, sampling effort needs to increase in order to make a decision at the prescribed power and significance. At some point, the sampling effort becomes so costly that judgement is reserved; i.e., no decision is made.

9.4 Management

9.4.1 Personnel

Trained and experienced biologists should be available to provide thorough evaluations, provide support for various activities, and serve as QC checks. They should have training and experience commensurate with the needs of the program. At least one staff member should be familiar with establishing a QA framework. QA programs should document personnel responsibilities and duties and clearly delineate project organization and lines of communication (USEPA 1995). A time line illustrating completion dates for major project milestones or other tasks can be a tremendously useful tool to track project organization and progress.

9.4.2 Resources

Laboratory facilities, adequate field equipment, supplies, and services should be in place and operationally consistent with the designed purposes of the program so that high-quality environmental data can be generated and processed in an efficient and cost-effective manner (USEPA 1992b). Adequate taxonomic references and scientific literature should be available to support laboratory work, data processing, and interpretation.

9.5 Operational Quality Control

Protocols should be developed for designing a data base and for screening, archiving, and documenting data. Data screening identifies questionable data based on expected values and obvious outliers. Screening is especially important if data are gathered from a variety of sources and the original investigators and data sheets are no longer available. The following text box defines the qualitative and quantitative data characteristics that are most often used to describe data quality.

Six qualitative and quantitative data characteristics usually employed to describe data quality:

1. *Precision* - The level of agreement among repeated measurements of the same characteristic.
2. *Accuracy* - The level of agreement between the true and the measured value, where the divergence between the two is referred to as bias.
3. *Representativeness* - The degree to which the collected data accurately reflect the true system or population.
4. *Completeness* - The amount of data collected compared to the amount expected under ideal conditions.
5. *Comparability* - The degree to which data from one source can be compared to other, similar sources.
6. *Measurability* - The degree to which measured data exceed the detection limits of the analytical methodologies employed; often a function of the sensitivity of instrumentation.

These measurement quality indicators require a priori consideration and definition before the data collection begins. Taken collectively, they provide a summary characterization of the data quality needed for a particular environmental decision. Duplication of approximately 10 percent of the total sampling effort is a common level for operational QC. Replication of samples at a randomly selected subset of field sites (usually, 10 % of the total number is considered appropriate) is used to estimate precision, and representativeness of the samples and the methods; splitting samples into subsamples can be used to check precision of the methodology, and reprocessing of finished samples is used to check accuracy of laboratory operations.

9.5.1 Field Operations

For the field operations aspect of an ecological study, the major QC elements are instrument calibration and maintenance, crew training and evaluation, field equipment, sample handling, and additional effort checks. The potential errors in field operations range from personnel deficiencies to equipment problems. Field notes are integral to the documentation of activities and can be used to help locate potential recording errors. Training is one of the most important QC elements for field operations. Establishment and maintenance of a voucher specimen collection should be considered for biological data. Transcription errors during data entry can be reduced with double data entry. Table 9-7 gives examples of QC elements for field and laboratory activities.

Table 9-7. Example QC elements for field and laboratory activities

Project Activity	QC Element	Evaluation Mechanism
Field Sampling	Replicated samples at 10 percent of sites by same field crew.	Calculate relative percent difference (RPD) of index value or individual metric score
	Replicated samples at one to two of total sites by different field crew using same methods.	Calculate RPDs as above; use to evaluate consistency and bias.
Physical Habitat Assessment (Qualitative)	Ensure appropriate training and experience of operators; multiple observers.	Resume or other documentation of experience; discuss and resolve differences in interpretation.
Physical Habitat Assessment (Quantitative)	Replicated measurements at 10 percent of sites.	Calculate RPDs between replicate measurements; compare to preestablished precision objectives.
Laboratory: Sample Sorting	Sample residue checked for missed specimens to estimate sorting efficiency; check completed by separate lab staff.	Calculate percent recovery; compare to preestablished goals.
Laboratory: Sample Tracking	Logbook with record of all sample information.	Not applicable.
Laboratory: Taxonomic Identification	Independent identification and/or verification by specialist; ensure appropriate and current taxonomic literature available; adequate training and experience in invertebrate identifications; reference collection; exchange selected samples/specimens between taxonomists.	Calculate percent error; compare to preestablished goals.
Data Management	Proofreading; accuracy of transcription.	All transcribed data entries compared by hand to previous form—handwritten raw data, previously computer-generated tables, or data reports.
Data Analysis	Hand-check of reduced data.	For computer-assisted data reduction, approximately 10 percent of reduced data recalculated by hand from raw data to ensure integrity of computer algorithm.
	Appropriate statistics; training.	Review by statistician or personnel with statistical training.

9.5.2 Laboratory Operations

The QC elements in laboratory operations include sorting and verification, taxonomy, duplicate processing, archival procedures, training, and data handling. Potential error sources associated with sample processing are best controlled by staff training. Controlling taxonomic error requires well-trained staff with expertise to verify identifications. Counting error and sorting efficiency are usually the most prominent error considerations; they can be controlled by training and by duplicate processing, sorting, and verification procedures. See Table 9-7 for examples for QC elements for laboratory activities.

9.5.3 Data Analysis

Errors can occur if inappropriate statistics are used to analyze the data. Undetected errors in the data base or programming can be disastrous to interpretation. Problems in

managing the data base can occur if steps are not taken to oversee the data handling, analysis, and summarization. The use of standardized computer software for data base management and data analysis can minimize errors associated with tabulation and statistical analysis. A final consideration is the possible misinterpretation of the findings. These potential errors are best controlled by qualified staff and adequate training.

9.5.4 Reporting

QC in reporting includes training, peer review, and the use of a technical editor and standard formats. The use of obscure language can often mislead the reader. Peer review and review by a technical editor are essential to the development of a sound scientific document.