Implementing the Action Plan

October 2001

Swan-Canning Cleanup Program

Derivation of Percentile Quality Criteria for the Swan-Canning Estuary, a binomial approach

SWAN RIVER TRUST

Acknowledgments

The authors would like to acknowledge the staff of the Swan River Trust and the Aquatic Sciences Branch of the Water and Rivers Commission for data collection, management and advice. Special thanks must go to Philipa Wittenoom and Emer O'Gara for their technical assistance, Dr Tom Rose and Malcolm Robb for their editorial comments and Robyn Paice and Virginia Shotter for their early work on this project. Lastly, a special thanks to Dr Bill Maar for his invaluable support and feedback regarding this work.

Reference details

The recommended reference for this publication is:

Donohue, R. and van Looij E. 2001, The Derivation of Percentile Qualtiy Criteria for the Swan-Canning Estuary; A Binomial Approach. Swan River Trust, Swan-Canning Cleanup Program Report SCCP 24.

ISBN: 0-7309-7562-2

ISSN: 1444-3279

Abbreviations and Symbols

n	= Number of random and independent samples
e	= Number of excursions from a percentile criteria in a set of n samples
e_{max}	= Maximum number of permissible excursions from a percentile criteria
p	= Probability associated with a one sided hypothesis test
x	= Excursion rate in tested ecosystem
$x_{\%ile}$	= Maximum allowed x defined nominally by a percentile
x_t	= Detectable threshold x (at 1-α) given n, the decision stance and the tested
n	= Number of random and independent samples
α	= Probability of Type I error
β	= Probability of Type II error
CI	= Confidence interval

2. Statistical concepts

The quality criteria developed for the Swan-Canning estuary are percentile criteria. Percentile ($_{\%ile}$) standards specify that quality should be no worse than a defined limit for more than a set percentage of time. For example, a $90_{\%ile}$ criteria specifies that quality can be worse than a criteria value but not for more than 10% of the time. The proportion of time that a tested system is greater than (or less than) a limit is known as the population excursion rate. For the $90_{\%ile}$ quality criteria, the maximum allowable excursion rate is 10% of the assessment period. If the rate of excursion were higher than 10% it would be concluded that the tested system is in breach of the criteria. For a $50_{\%ile}$, $60_{\%ile}$ and $80_{\%ile}$ quality criteria, the maximum allowable excursion rate is 50, 40 and 20%, respectively.

The period in which ecosystem quality is above or below a threshold level is not usually known for an aquatic systems. In reality the true rate of excursion from a limit can never be known with absolute certainty — it can only be estimated using monitoring data. Even if data are collected specifically to identify the excursion rate from a criteria, because the estimate is based on a limited set of data its accuracy will always be uncertain. When the results of monitoring are to underpin management decisions it is important that allowances are made for sample error and the level of uncertainty.

2.1 Compliance with percentile quality criteria

The number of sample excursions from a criteria will depend on 1) the number of samples collected and 2) the (unknown) rate of excursion in the estuarine system. Assume 60 samples are collected from an estuary to test compliance with a nitrate criteria of 45 µg/L. If the rate of excursion in the estuary from 45 µg/L nitrate were 10%, 50% or 80% of the assessment period, the number of samples with more than 45 µg/L would be somewhere around 6, 30 and 48, respectively (although the exact number that could be collected in each case will vary).

Since the number of sample excursions reflects the rate of excursion, the number of excursions in a set of n samples can be used as the weight of evidence supporting the compliance decision. For example, if sample error were ignored and the results of monitoring taken simply on face value, if more than 6 high samples were collected it could be concluded that the excursion rate is higher than 10% and therefore the quality of the tested estuary is worse than that specified by a $90_{\%ile}$ criteria. That is the tested system would be found to breach its quality condition if 7 or more samples were found to contain more than 45 µg/L nitrate.

However, the actual number of high samples that are collected is, to some extent, a matter of chance. The number of excursions in 60 samples will vary from one trial to the next even if the excursion rate is constant. If the excursion rate were 10%, it would not be unusual if only 5 high samples were collected or even as few as 4. Equally it would be no surprise if 7 or 8 high samples were to be collected in 60 samples. This random variation in the number of sample excursions raises the question: how many high samples should be allowed before it is concluded that the excursion *rate* is inconsistent with that nominated by the percentile criteria?

2.2 The probability of excursion

Variation in the number of high samples found in a set of n samples can be predicted (Miller and Ellis 1986; Ellis 1989). Provided sampling is random and independent, the variation in sample excursions from a set of n samples will always follow a binomial distribution (regardless of the distribution of the underlying process). The binomial distribution describes the behaviour of any system in which there are only two possible outcomes (such as heads or tails, on or off, male or female, higher than or less than). The probability of getting '*e*' excursions from a set of n random samples is:

breach of the criteria. This error is known as a Type II error (denoted as β) and is known as the "consumers" risk because its consequences are borne by the water user or receiving environment (McBride and Ellis 2000).

samples may be collected and an assessor will still conclude compliance. But with 10 high samples of 60 the actual excursion rate could be as high as 26.6%. Therefore in adopting this approach the program designer has accepted that quality must be worse than the criteria for *at least* 26.6% of the time before it will be detected.

The blue curve in Figure 2 is the power curve for the precautionary fail-safe decision rule "$e > 2$ = breach". With the fail-safe approach notice that it is the Type II error that is controlled at 5% for marginal breach (10.1% in the example) and the excursion rate where α is 5% is consequential. With fail-safe, if the rate of excursion in the tested system is 10.1% or more it will be found to breach the quality criteria on 95% of occasions.

The two power curves illustrate the trade off that must be made when designing a compliance scheme in this case for a $90_{\%ile}$ criteria. With the fail-safe approach there is an almost 95% probability of a Type I error when x is exactly 10%. With benefit-of-doubt x must degrade some way beyond that specified by the percentile ($x_{\%ile}$) before a breach will be detected. On the other hand, in the attempt to detect a very small deviation from the nominal $x_{\%ile}$ (fail-safe), an assessor can expect a high number of false alarms. The relative costs of both approaches should be weighed carefully before settling on a decision stance.

3.2 The allocation of risk

The incidence of risk depends on the hypothesis being tested (McBride and Ellis 2000). If it is decided that the Type I error risk will be kept small (the benefit-of-doubt scheme), then the tested hypothesis must be one of compliance. When the prior assumption is compliance, the Type I error is the suppliers risk and the Type II error risk is the consumers risk. If it is decided that the probability of a Type II error is to be controlled (fail-safe), the hypothesis tested will be breach. If the tested hypothesis is breach the incidence of risk is reversed — the Type I error risk becomes the consumers risk and the Type II error risk is the suppliers risk (McBride and Ellis 2000).

3.3 Estimating the current rate of excursion

Suppose that a single sample is collected from a lake every week for a period of one year. Of the 52 samples collected in the first year of monitoring, 12 are found to contain more than 100 µg/L of nitrate nitrogen (NO_3-N). The question is: given $e = 12$ from a set of 52 samples, what is the best estimate of the excursion rate above 100 µg/L NO_3-N in the lake? This question can be answered through the use of confidence intervals around proportions. A confidence interval (CI) is a range that contains the true rate of excursion x. A 90% CI for example is the range in which it is 90% certain x lies. It is not know where the true x lies it is only known that it lies somewhere within the range.

The low end of the CI (denoted as P_{low}) is known as the optimistic end of the CI because it assumes that the number of sample excursions was higher than could be expected given x. To calculate P_{low}, a value must be found for x that is so low that the probability of getting as many as 12 high samples out of 52 is just 5% (the 100-2α CI). The value of x that satisfies this condition is found through a process of trial and error. The data suggest the concentration of NO_3-N in the lake was above 100 µg/L for somewhere in the vicinity of 23% of the year.

To start try a value for x that is lower than the rate of excursion suggested by the data set. First try say x = 20%. From Equation (1), if x were 0.2, the p of getting $e \geq 12$ is 34%. Since this is much larger than 5%, try a lower x, this time x = 15%. Again from (1), with x = 0.15, there is an only an 8% chance of getting $e \geq 12$ from 52 weekly samples. This is close but still too high so try x = 14%. With a system in which x = 14%, there is a 5.3 % chance of getting $e \geq 12$ out of 52 samples. This is probably close enough but to be exact, if the concentration of NO_3-N in the samples was above 100 µg/L for 13.9% of the year, there would be only a

4. The confidence interval method

Uncertainty due to sample error can be integrated into the decision making process using the CI method (Ellis 1989). With the CI method, the compliance decision is made with reference to the location of the upper and lower ends of the CI relative to the quality criteria (denoted as x_{max}).

Figure 4 shows how the CI method works. With Case A and D in Figure 4 there is little doubt associated with the conclusion reached — the entire CI are above x_{max}. With Cases B and C, however the decision is more difficult.

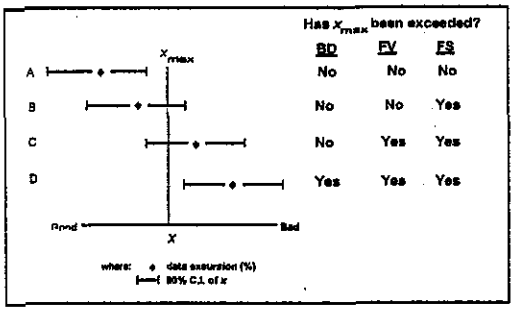

Figure 4. Using the CI method to conclude if the $x_{\%ile}$ has been exceeded. The plot shows the conclusions reached for the benefit-of-doubt (BD), face-value (FV) and fail-safe (FS) approaches (adapted from Ellis 1989)

In Case B, while the data excursion rate (the diamond-shaped symbol) is below x_{max}, the upper end of the CI extends beyond it. It is therefore conceivable that quality in the tested system is worse than x_{max}. In adopting the fail-safe position an assessor would conclude breach but, with the benefit-of-doubt approach, compliance would be concluded.

With Case C, the excursion rate of the data set is above x_{max} and on face-value an assessor would conclude breach. However since the lower end of the CI extends below x_{max}, it is possible that x in the tested system was better than the quality objective, so the benefit-of-doubt approach produces a verdict of compliance.

4.1 Identifying change in excursion rates

An increase in x indicates that there has been an increase in the time that quality is worse than a criterion. Even a relatively small change in excursion can be used to warn a management agency of a possible degradation in quality. Decreases in the proportion of time (x) that a tested system is worse than a criteria value can be used to signal improvements towards meeting management objectives for degraded systems.

The NWQMS Guidelines on Monitoring and Reporting (2000 Draft) recommend that a staged system of warnings be used, each stage representing an upgrade in the level of alert. The stages are: 1) while a test system is below a threshold quality no action is required; 2) quality equal to the threshold quality (i.e. $x = x_{\%ile}$) results in a warning that an investigation may be required; and 3) quality becoming greater than the threshold quality means that degradation is confirmed and the next level of investigation is necessary (there are no recommendations for critical stages of improving quality).

In the current format (Draft NWQMS 2000) the staged system of warnings of degradation do not make allowance for sampling error (i.e. they have recommended the face-value approach). However, the staged warning system can be used with either the fail-safe or benefit-of-doubt approaches through the CI method.

4.1.1 Detecting increases in excursion

Figure 5 shows a time series chart showing how the CI method would alert a management agency to a degradation in quality. The example is a $90_{\%ile}$ criteria for chl-*a* concentration. The criteria value is 3.35 µg/L. Therefore the aim of the Swan-Canning Cleanup Program is

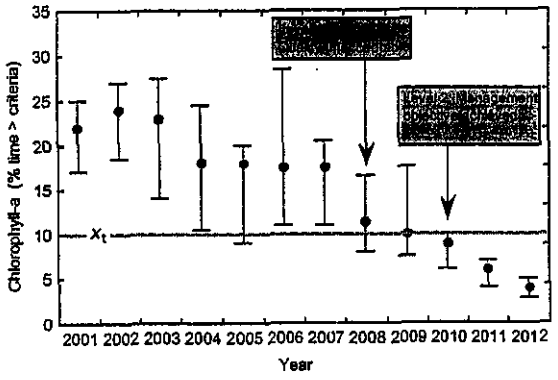

Figure 6. How the CI method can be used to identify the achievement of a quality objective for a degraded system

In the year 2001 the entire CI is well above the threshold at x_t and remains above it until 2005. In 2005 the low end of the CI falls below the line at x_t but most of the interval is still above the line. Since the true rate of excursion in the tested system may reasonably be above the line it is concluded that the management quality objective has not yet been achieved. Subsequently the CI again moves well above the line x_t.

From 2007 the excursion rate appears to be falling but the decision that the management objective has been met is delayed until the upper end of the CI falls below x_t. In the example this result occurs in the year 2010. The rule that governs notification that the management program has achieved its goal is: $e < 3$ = *management objective achieved.* In the Figure 6 example, $e < 3$ was recorded for the first time in the year 2010.

4.2 The threshold rate of excursion X_t

Once the percentile and n have been chosen, the location of x_t (in Figures 5 and 6) depends on whether an assessor adopts a benefit-of-doubt or fail-safe approach.

If an assessor adopts a benefit-of-doubt approach to decision making, x_t is determined by the requirement to control the Type I error. For n = 60 and a $90_{\%ile}$ criteria, the benefit-of-doubt decision rule is: "$e > 10$ = breach". With this compliance there is a 95% chance that a marginal breach would not be detected. Remember, that the chance of a Type II error falls to 5% only if x is at least 26.6% (Figure 2). The smallest rate of excursion (x_t) that can be detected with 95% certainty with the compliance rule "$e > 10$ = breach" is 26.6%. Therefore for this example the threshold excursion rate x_t would be 26.6%. The issue of a Level 3 warning would be made with reference to the lower end of the CI and the x_t threshold defined by the Level 2 warning (Figure 5).

If the fail-safe position is adopted in detecting degradation, the location x_t is determined by the requirement to control the Type II error (usually 5%) using the fail-safe compliance rule, which is "$e > 2$ = breach" for a $90_{\%ile}$ criteria and n = 60. With this scheme the threshold x_t would be 10.1% (Figure 2).

The situation for when the tested hypothesis posits breach (and not compliance) is depicted in Figure 6. With breach as the a priori position the allocation of risk is reversed — the supplier's risk is the Type II error probability (McBride and Ellis 1989). This means that the decision that the management objective is achieved will be made with reference to the threshold at $\beta = 5\%$ at marginal breach. As Figure 2 shows this is "$e \leq 2$ = compliance" and x_t is located at 10.1%.

Table 2. Quality criteria for total nitrogen (TN), total phosphorous (TP), chlorophyll-*a* (chl-*a*) and dissolved oxygen (DO) for the Swan-Canning estuary

Estuary	Compliance Region*	Depth	Variable	Criteria	$x_{\%ile}$ (% time)	x_t	Current x (90% CI)
Swan	L	S	TN (µg/L)	509.0	10	10.1	4 — 10
Swan	M	S	TN (µg/L)	790.0	10	26.6	28 — 35
Swan	U	S	TN (µg/L)	1009.0	10	26.6	27 — 35
Canning	M	S	TN (µg/L)	790.0	10	26.6	27 — 46
Canning	U	S	TN (µg/L)	1330.0	10	26.6	26 — 45
Swan	L	S	TP (µg/L)	58.0	10	10.1	4 — 11
Swan	M	S	TP (µg/L)	110.0	10	26.6	28 — 35
Swan	U	S	TP (µg/L)	119.0	10	26.6	27 — 35
Canning	M	S	TP (µg/L)	190.0	10	26.6	27 — 46
Canning	U	S	TP (µg/L)	300.0	10	26.6	27 — 46
Swan	L	S	DO (% sat)	82.1	5	5.6	2 — 5
Swan	M	S	DO (% sat)	75.1	5	18.8	19 — 25
Swan	U	S	DO (% sat)	81.2	5	18.8	19 — 26
Canning	M	S	DO (% sat)	49.1	5	18.8	19 — 29
Canning	U	S	DO (% sat)	15.4	5	18.8	19 — 29
Swan	L	B	DO (% sat)	33.7	5	5.6	3 – 6
Swan	M	B	DO (% sat)	32.9	5	18.8	19 – 27
Swan	U	B	DO (% sat)	12.1	5	18.8	19 – 33
Canning	M	B	DO (% sat)	36.4	5	18.8	19 – 35
Swan	L	S	chl-*a* (µg/L)	3.55	10	10.1	4 — 11
Swan	M	S	chl-*a* (µg/L)	8.75	10	26.6	27 — 35
Swan	U	S	chl-*a* (µg/L)	19.98	10	26.6	28 — 36
Canning	M	S	chl-*a* (µg/L)	11.67	10	26.6	30 — 49
Canning	U	S	chl-*a* (µg/L)	39.00	10	26.6	28 — 47

* L = Lower estuary, M = Middle estuary, U = Upper estuary

** Current rate of excursion in the estuary basins estimated from monitroing data 1996 - 1998

Table 3. Chlorophyll-*a* concentration (µg/L) and oxygen (% saturation) in surface waters of the lower Swan-Canning and upper Canning estuary, respectively. Data were collected weekly between January and May 1996 to 1998. The table contains only the 59 highest ranked chlorophyll-*a* values of the 192 samples collected from the lower Swan-Canning, and the 62 lowest ranked dissolved oxygen measurements from the 163 samples collected in the upper Canning.

Chlorophyll in the lower Swan estuary				Oxygen in the upper Canning estuary			
Rank*	Value	Rank	Value	Rank*	Value	Rank	Value
1	16.00	32	3.00	1	1.70	32	12.40
1	16.00	32	3.00	2	2.10	33	12.80
3	7.20	34	2.90	3	2.40	34	13.10
4	5.60	34	2.90	4	2.50	35	13.20
4	5.60	34	2.90	5	2.70	36	13.30
6	5.20	37	2.80	6	3.60	36	13.30
7	4.70	37	2.80	6	3.60	38	13.50
8	4.40	37	2.80	6	3.60	39	14.40
9	4.30	37	2.80	9	3.70	40	14.50
10	3.90	41	2.70	10	3.90	41	15.40
11	3.80	41	2.70	10	3.90	42	15.60
12	3.70	41	2.70	12	4.00	43	15.70
13	3.60	41	2.70	12	4.00	44	16.40
13	3.60	45	2.60	14	4.20	45	17.60
15	3.50	45	2.60	14	4.20	46	18.30
15	3.50	45	2.60	16	4.40	47	18.80
17	3.40	45	2.60	17	4.50	48	19.00
17	3.40	45	2.60	18	4.70	49	20.00
17	3.40	50	2.50	19	4.90	50	20.80
17	3.40	50	2.50	20	5.00	51	21.60
17	3.40	50	2.50	21	5.10	52	22.50
17	3.40	50	2.50	22	6.10	53	22.60
23	3.30	54	2.40	23	8.30	54	23.60
23	3.30	55	2.30	24	8.90	55	24.90
25	3.20	55	2.30	25	9.10	56	25.40
25	3.20	55	2.30	26	9.30	57	25.80
27	3.10	55	2.30	27	10.00	58	26.70
27	3.10	59	2.20	28	10.20	59	27.40
27	3.10	59	2.20	29	11.30	60	27.60
27	3.10	59	2.20	30	11.70	61	28.40
27	3.10	59	2.20	31	11.80	62	30.80

The ranking method used here gives duplicate numbers the same rank. However, the presence of duplicate numbers affects the ranks of subsequent numbers. For example, in a list of integers, if the number 10 appears twice and has a rank of 5, then 11 would have a rank of 7 (no number would have a rank of 6)

estuary. From Table 2, the 41st ranked DO saturation value is 15.4 (%sat). This criterion is a provisional quality objective for the Swan-Canning Cleanup program indicating only that the frequency of occurrence of very low DO during the assessment period has decreased.

5.3.3 The decision rule

The rule that signals the achievement of the quality objective for DO in the upper Canning is:

LEVEL 1 [redacted]

If more than 1 of 60 measurements of DO are below 15.4% saturation.

LEVEL 2 [redacted]

If less than 2 of 60 measurements of DO are below 15.4% saturation.

References

ANZECC (2000). *Draft June 2000 Australian and New Zealand Guidelines for Fresh and Marine Water Quality – Volume 1: The Guidelines*

Ellis, J.C. (1989). *Handbook on The Design and Interpretation of Monitoring Programs.* Water Research Centre, Medmenham, UK.

Gilbert, R.O. (1987). *Statistical Methods for Environmental Pollution Monitoring.* Van Nostrand Reiinhold, New York.

Hart, B., Maher, W., and Lawrence, I. (1999). New generation quality guidelines for ecosystem protection. *Freshwater Biology* **41**: 347 – 359.

McBride, G. and Ellis, J.C. (2000). Confidence of compliance: A Bayesian approach for percentile standards. Accepted by *Water Research* 8th May 2000.

Miller, D.G. and Ellis, J/.C. (1986). Derivation and monitoring of consents. *Water Pollution Control* 249 - 258

Swan River Trust. (1999). *Swan-Canning Cleanup Program Action Plan.* May 1999.

Ward, R., Loftis, J. and McBride, G. (1990). *Design of Water Quality Monitoring Systems.* Van Nostrand Reinhold, New York.

as a consequence of controlling the Type I error. For the chl-*a* example, the decision threshold is 26.6%. The power curve for the rule $e > 10$ = 'breach' also defines the decision threshold (the line x_t in Figure 9) to upgrade the level of protection or surveillance. In the example, the decision to upgrade the level of warning of degradation is made with reference to the location of the lower end of the CI (P_{low}) compared to x_t at 26.6%. For the example depicted, P_{low} will be to the right of x_t when e = 23 of n = 60.

If the fail-safe position had been adopted, the β error risk would have been controlled at 5% and the decision threshold in Plot A (the red line) would have been at 10.1% and not 26.6%. The first decision "*warning, an investigation may be necessary*" would be made when $e > 2$, and the second decision to upgrade the response to "*next level of investigation*" when $e > 10$. The quality criteria would have been selected so that the upper end of the CI was below 1.4%, where $\alpha = 0.05$ for the "$e > 2$ = breach" rule (see Figure 2).

Arranging breach

Plot B in Figure 9 shows the situation when the tested hypothesis is breach. The example is for DO in the upper basin of the Canning estuary. This criteria establishes a condition on the $5_{\%ile}$ oxygen saturation (DO stresses ecosystems at low levels). Since in this case the tested hypothesis posits breach, the Type II error is controlled to a maximum of 5% and the rate of excursion at α = 5% is consequential. Note also that with a 'breach' hypothesis the incidence of risk switches so that the Type II error is now the suppliers risk.

To arrange it so that the Type II error rate is truly 5% at most, a quality criteria was selected so that current rate of excursion was to the right of the threshold line at 18.8% (for the decision rule "$e > 6$ = breach"). The decision rule that restricts β = 5% is '$e < 2$ = compliance'. This rule places x_t at 7.7% (Figure 9). If e = 1 from n = 60, P^{high} of x is 7.6% and it will be concluded (with 95% confidence) that the tested system is no longer in breach and that there has been an improvement in quality

A Brief Overview of Statistics

Statistics is the science used to infer truth about the real world using only limited numbers of samples. It can be used to make accurate decisions under ~~circumstances of~~ uncertainty. The collection of all possible measurements in space and time from an object or objects of interest is called the *population*. Rarely can all measurements from a population be taken, however. Population size may be finite, but more often it is too large to measure in its entirety.

An example of a population is the concentrations of dissolved oxygen at all points in the lower San Joaquin River over a two year period. An investigator researching the health of the River cannot possibly know oxygen concentrations at all points in the water body. Instead, she would commonly rely on *samples* of oxygen concentrations taken appropriately and in suitable numbers from various points and at various times in the River. Statistical analysis could then be used to relate sample findings to conditions in the population as a whole.

Statistics is ~~soundly~~ based in probability theory. Both are rational, quantitative disciplines based on logic, numeric evaluation, and mathematics. Probability at its simplest is concerned with the likelihood of a particular event occurring. The chance of such an event occurring ranges from zero (0%; there is no chance the event will occur), to one (100%; the event will definitely occur).[1]

Population Parameters

To choose the right statistical tests in order to reach valid conclusions, researchers need to first establish certain characteristics of the population in question. Populations are frequently characterized by their *central tendency* and by their frequency *distribution* around some central point. Various *parameters* (numeric values) are used to determine these characteristics; familiar parameters associated with central

1 For example, the likelihood of rolling a "1" is one out of six or 1/6 (approximately 0.16667, or 17%), for a fair die ~~on a level table.~~ The probability of rolling six "1's" in a row, one after the other, is 1/6 x 1/6 x 1/6 x 1/6 x 1/6 x 1/6 (= [1/6]6), or 1/46,656 (0.00214%)—an extremely unlikely event. On the other hand, the likelihood of rolling either a "1," a "2," a "3," a "4," a "5," or a "6" is 1/6 +1/6 +1/6 +1/6 +1/6 +1/6 = 1 (100%; a certainty).

~~These examples illustrate basic laws of probability.~~ The probability of independent events can be calculated by multiplying their individual probabilities. However, the total likelihood of all mutually exclusive possible results from taking a single action is determined by adding the individual probabilities, which if all possible events are accounted for should add to one (100%). Along with mathematics, probability theory forms the basis for calculating statistical distributions and associated tests.

tendency include the range, mean, variance, and standard deviation of the population. Although these values are extremely helpful in understanding the nature of a population of measurements, they can seldom be known with certainty. Unless a population of measurements is very small, researchers cannot take all possible measurements necessary to determine population parameters with 100 percent assurance. Instead, population parameters are estimated from *samples*, small subsets from the population of all possible measurements. The challenge then becomes how to sample, analyze, and interpret findings in order to estimate population parameters with reasonable assurance and to make valid decisions based on the population of measurements.

Types of Statistical Tests

Statistics combines probability theory and mathematics in developing statistical tests intended to quantify the accuracy and applicability of numeric sampled values, or *data*. Commonly used statistical tests fall into two broad categories: parametric and non-parametric tests.

Parametric tests use relatively small amounts of data to calculate sample statistics, that are in turn estimates of true population parameters. In contrast, nonparametric tests often transform initial sample results (e.g., from raw numeric values to rankings, or from numeric data to nominal information) and draw conclusions about the population without estimating population parameters directly. Both types of tests offer advantages and disadvantages. To meaningfully compare the two types of tests various topics related to statistics—i.e., error rates, assumptions, distributions of data, and desirable attributes--must be reviewed.

Statistical Error

Could be put in Hypothesis Issue

Statistics is often used to bolster decision-making when conditions are uncertain. One common statistical procedure is *hypothesis testing*. A researcher starts with a statement of the status quo called the *null* hypothesis. For 303(d) listing/delisting an example might be a mathematical expression implying that a lake's water quality is acceptable for a particular substance (i.e., the concentration of the pollutant in the lake is below the threshold criterion for that pollutant) and that it should therefore not be placed on the section 303(d) list. The alternative hypothesis would then be that the water body is polluted by that substance (i.e., the criterion is exceeded some critical percentage of the time) and it should be listed. For statistical purposes, the two possible

hypotheses may be reversed--"the water is polluted" could be the null hypothesis with "the water is not polluted" serving as the alternative hypothesis. In either case, the lake's water would be sampled, the samples tested for various pollutants, and the results compared against existing water quality objectives. Decision-makers could then reject the null hypothesis (whichever one is used), or not, depending on what the results showed.

Critical to this process is the fact that two types of error may occur in hypothesis testing (see table, below). The first type of error, called a *Type I* (false positive) error, occurs when a true null hypothesis is erroneously rejected. If the null hypothesis is that a water body is not polluted, then concluding that it is polluted and listing it when it truly is not polluted is an error of the first type. Suppose it is decided that a water body should be designated as polluted when it violates a water quality objective more than ten percent of the time. Suppose also that a river is actually in violation on the average only one percent of the time. Despite this, and even with adequate sampling design and laboratory analysis, occasionally over 10 percent of samples may exceed a criterion and result in a Type I error by investigators.

Types of Statistical Error
(With H_0 = water body is meeting water quality standards)

Decision	**H_0 True (standards met)**	**H_0 False (standards not met)**
Reject H_0 (list)	**Type I (false positive) Error** (list when inappropriate to)	**Correct Decision**
Do not reject H_0 (do not list)	**Correct Decision**	**Type II (false negative) Error** (do not list when appropriate to)

Statisticians signify the chance of a Type I error by the Greek letter "alpha" (α). The value of alpha ranges from zero to 100 percent, and can be determined/controlled accurately for the various statistical tests, often because tests are designed such that the alpha value is pre-selected by the investigator, who then has almost complete control over what maximum alpha level will be tolerated. In addition, researchers strive through proper sample design and analytical procedures to keep the minimum chance of making a Type I error as low as possible. The likelihood of not making a Type I error, one minus alpha ($1 - \alpha$), is known as the *confidence*. The higher the confidence in a test result, and therefore the lower the alpha rate, the less likely a conclusion based on the sample

results will result in a Type I error. If the initial premise to be tested, the null hypothesis, is that a water body is achieving water quality standards and should not be listed, and if this premise is indeed true, then a statistical analysis with a low alpha rate (i.e., high confidence) improves the chance of correctly not listing the water body.

The second type of error, a *Type II* (false negative) error, occurs when a false or untrue null hypothesis is erroneously not rejected. For example, if the null hypothesis is that the water is not polluted, then concluding that a particular water body is not polluted and leaving it unlisted when it is indeed polluted is an error of the second type.

The Greek letter "beta" (β) is used to signify the chance of making a Type II error. The chance of avoiding a Type II error, one minus beta ($1 - \beta$), is known as the *power* of a test. The greater the power, the less the likelihood of making a Type II error. For example, if the null hypothesis is that a water body is achieving water quality standards and should not be listed, and the water body happens to be impaired; statistical results with a low beta rate (i.e., high power value) will mean that the chance of correctly listing the polluted water body is high.

As stated above, most statistical tests are designed so that alpha is directly included in the statistical analysis by the investigator. However, unlike alpha beta is not normally a user-controlled variable in statistical tests and associated software. Nevertheless, both power and confidence tend to improve (i.e., α and β decrease) with increased sample sizes.

Statistical Test Assumptions

Correctly applying statistical tests often requires that various assumptions of the data and data collection have been met. For example, many tests require that data collection (sampling) be randomly performed. Furthermore, sample results must often be independent of one another. Some parametric tests require that sample data have originated from a population of normally distributed or continuous measurements. Some tests require that values not be ratios or percentages. In general, statistical tests may function when assumptions are not met, but their results may be less powerful or reliable.

Distributions of Data

When numeric information is plotted for frequency, statisticians find that data falls into discernable types of *frequency distributions*. For example, many types of unrelated data are *normally* distributed resulting in the classic bell-shaped, symmetric curve around a single arithmetic mean/average (also the median [frequency mid-point] and mode [highest point]of the normal distribution). Many types of phenomena, but not all, when sampled in adequate numbers appear to originate from normally distributed populations of data. Other types of distributions include the bivariate normal, Gaussian, hypergeometric, binomial, Poisson, and lognormal. Knowledge of the type of population distribution that sample data originate with is important, since statistical tests can be more reliable when applied to data from particular distributions.

Desirable Statistical Attributes

Statistics are estimates of population parameters. A sample mean is an estimator for the true population mean, which by the way is probably unknowable. Other statistics, such as a sample median (the value for which 50% of sample values are below and 50% above), may also be estimators of a population mean. In cases where there is more than one choice, which statistic is most reliable for determining information about a population? According to Zar (1999), a reliable statistic/estimator is:

unbiased. Given an infinite number of samples, the average of the statistic's values from all samples would equal the population parameter being estimated.

precise. The statistic from any one sample is close to the actual population parameter being estimated.

consistent. As sample size increases, the sample statistic grows closer and closer to the population parameter being estimated.

Parametric versus Nonparametric Tests

Parametric statistical procedures result in sample statistics that estimate population parameters. For example, the sample mean may be an effective estimate of the true, but unknown, population mean. The range detected in a sample may reliably stand in for the true population range. Measures of variability such as sample variance and standard deviation may closely approximate the true variability around the population mean. Parametric statistical tests often include, or are mathematically-related to, these descriptive

statistical characteristics (variance, standard deviation). An important distinction is that parametric procedures use actual sample data.[2]

Nonparametric procedures are not based on, derived from, or associated with population parameters such as the mean or variance. They are nonetheless mathematically valid and well-studied. Many of the better known nonparametric tests substitute *ranks* in place of the actual sample values. The data are ranked by magnitude; and these numeric ranks rather than the original data are used within the statistical computations.

For one nonparametric procedure, the exact binomial test, numeric sample data are first converted into nominal information—i.e., either a "yes" or "no," "on" or "off," or other form of dichotomous result. Counts or proportions of the two types of results are then compared to tallies or frequencies derived from the well-understood binomial distribution for the sample size in question and other variables. In many basic nonparametric tests, the original, numeric values are not used in calculations. While this results in a loss of information, nonparametric procedures free the instigator from initial worries about population distributions and statistical assumptions. Nonparametric statistics are sometimes called distribution-less statistics.

Frequently, both parametric and nonparametric tests are available to perform the same or similar basic analyses. In general, parametric tests are more powerful when the assumptions of the tests are met. However, when test assumptions are violated significantly—for example, when the distribution of sample data are heavily skewed as occurs when samples sizes are small--the inherent power of nonparametric tests does not decrease as significantly as with parametric tests. It can be difficult to determine whether parametric test assumptions are being met. Again, nonparametric tests reove these concerns, albeit at a slight loss of statistical power.

Many parametric tests also tend to be robust (i.e., retain accuracy and precision) against small to moderate departures from tests assumptions (e.g., normality, equal variances among samples, etc.). But when departures are great, nonparametric tests can give more accurate or powerful results. Nonparametric test results are often easier to calculate than their parametric counterparts. Also, nonparametric tests often transform data (e.g., from raw numbers

2 Sample data may be transformed in order to normalize the sample distribution of values (i.e., make their frequencies approximate a normal distribution).

to rankings, or from numeric to nominal information), therefore, they can perform reliably with data from various types of distributions.