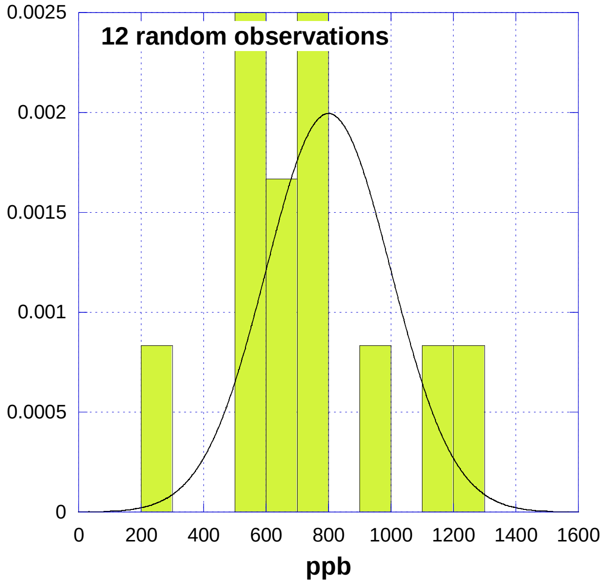
National Objectives Framework

Statistical considerations for design and assessment

Prepared for Ministry for the Environment

September 2016



Prepared by: Graham McBride

For any information regarding this report please contact:

Graham McBride
Scientist
Aquatic Pollution
+64-7-856 1726
Graham.McBride@niwa.co.nz

National Institute of Water & Atmospheric Research Ltd
PO Box 11115
Hamilton 3251

Phone +64 7 856 7026

NIWA CLIENT REPORT No: Ham16022
Report date: September 2016
NIWA Project: MFE16203

Quality Assurance Statement		
	Reviewed by:	Dr Neale Hudson
AB	Formatting checked by:	Alison Bartley
	Approved for release by:	Dr Rupert Craggs

Whilst NIWA has used all reasonable endeavours to ensure that the information contained in this document is accurate, NIWA does not give any express or implied warranty as to the completeness of the information contained herein, or that it will be suitable for any purpose(s) other than those specifically contemplated during the Project or agreed by NIWA and the Client.

Contents

Tables

Figures

Executive summary

The Ministry for the Environment has requested advice on a number of statistical issues pertinent to the 2014 National Policy Statement for Freshwater Management, having particular regard to its National Objectives Framework ('NOF'). Three broad topics are to be addressed: Interpreting current NOF statistics; Guiding a consistent approach to future attributes; Assessing the human health value.

Rather than immediately addressing these topics, this report first presents some fundamental statistical issues, and an analysis of the statistical and sampling issues raised by the current expression of 'Numeric Attribute States'. Foremost among these is the issue of sampling variability, sometimes denoted as 'statistical sampling error'. In particular, this means that we never know the true Attribute State (as a percentile of time), because we only ever have an estimate of that and such an estimate is influenced by the variability components we happen to capture in our samples. Sometimes it will be higher than the true value, sometimes lower. It will seldom if ever be the true value. Hence the proposed motto: "Always be wary of the influence of 'statistical sampling error'".

It is noted that for the most part the NOF is silent on the burden-of-proof that underlies the various percentile requirements, seemingly because they have been regarded as percentiles *of samples* rather than percentiles *of time*. If the latter is intended then there will need to be a direct consideration of the burden-of-proof and misclassification error risks when assessing Numeric Attribute States. The implications for sampling effort are detailed. It is noted that adopting a precautionary approach generally considerably increases the needed number of samples (proof-of-safety is more onerous than proof-of-hazard), as indicated in the look-up tables presented.

It is noted that the primary contact recreation human health Numeric Attribute State is already based on a precautionary approach to the burden of proof.

The propensity for "State Switching" (e.g., inferring states A-B-A-B-B in five successive years when in fact the waterbody was always in State B) has been analysed using a set of Monte Carlo numerical experiments and both normal and lognormal distributions. Under annual assessments, using only that year's data, the degree of switching seems unacceptably high. Instead it is suggested that a five-year assessment period be adopted with rolling annual assessment frequency.

A recently-developed (and implemented) direction-of-trend assessment procedure is recommended for progress assessment for the various attributes, after a few years of data have been analysed.

For the secondary contact in the human health value it is seems largely unnecessary to restrict sampling to seasons and lower flows, with one possible exception. That is, consideration should be given to (if possible) sampling for *E. coli* only on the rising limb of a flood hydrograph. For primary contact there should be sampling stratification based on season and on flow (when conditions may be unsuitable for swimming). The identification of what elevated flows and seasons should be so-treated will vary from location to location.

Consistency of approach should be aimed for with future Attributes (or revisions of the current set), but may not always be achievable. Issues that may arise with a percentage change approach (from a reference state or an upstream state) for sediment attributes will need careful consideration in light of the findings in this study.

Recommendations are made for future work, including:

(i) Conducting annual progress assessment for median thresholds using a rolling or (rarely) an adjacent window of at least three year's data, preferably five;

(ii) Using single-year adjacent assessments for high percentiles in cases where rare events may occur (e.g., lake hypolimnion hypoxia, *E. coli* spikes resulting from a WWTP failure);

(iii) Identifying appropriate burdens-of-proof;

(iv) Stating rules for inferring percentile attainment using simple "look-up" tables;

(v) Using two-one sided interval tests for state-switching considerations, such that a possible outcome is "U" (undecided);

(vi) Examining the feasibility of detecting a percentage change (for future attributes such as sediment) where a percentage change from some reference value is contemplated, given the potential for statistical sampling error to frustrate the ability to detect it;

(vii) Directing sampling for primary contact recreation toward times when these recreational activities are occurring;

(viii) Making some changes to the wording of Numeric Attribute States' assessment metrics to clarify their intention and scope.

Finally a three-step decision template is recommended, covering:

a) Choosing a time period and assessment regime;

b) Deciding on the burden-of-proof;

c) Choosing a comparison reference?

This template is applied in indicative fashion to total phosphorus in lakes; *E. coli* in rivers, and to a possible future attribute: visual clarity in rivers.

1 The Ministry's brief

1.1 Topics to be addressed

The Ministry for the Environment has contracted NIWA (via the author) to report on three broad issues concerning the NPS-FM National Objectives Framework (NOF), NZ Government (2014). These are quoted verbatim below (italicised text).

1.1.1 Interpreting current NOF attribute statistics

Guidance on how to interpret the current NOF attribute state sample statistics, given the problem of sample statistic imprecision. The issues to be addressed are the number of samples and sampling duration needed to characterise the state of a managed environment, and how the choices made when deciding a sampling regime can affect what the current attribute state of an environment is assessed as. Guidance will provide practical approaches available for councils to assess current state for the range of attributes. It is expected this will include advice concerning the intent of NOF attributes (i.e., that these define objectives as opposed to standards), and any implications imprecision has for the estimated state to be interpreted. The guidance should also include a tool (such as a look-up table) which can be used to determine an appropriate sampling regime given the numeric attribute percentile, necessary characteristics of the data or environment being sampled, and the appetite for uncertainty.

1.1.2 Consistent approach to future attributes

Advice for how the Ministry could define sample statistics in a consistent way for future attributes. This should include a review of approaches used for existing attributes, and a recommendation in the interest of improving consistency and clarity. The Contractor shall include guidance on how to specify "annual" or other such statistics, percentiles, and minimum sample durations and number of samples. In preparing this advice, the Contractor will consult with the coordinator of the ANZECC Guideline Revision (Dr Chris Humphries of the Australian Department of the Environment) and Dr Brent Henderson (a statistician at CSIRO) who are currently preparing guidance on sampling statistics. The Contractor will endeavour to align the approaches where applicable. To extent the approaches do not align, the Contractor shall explain the reasons for non-alignment. Reviewing these Attribute tables for inconsistencies and potential future improvements;

1.1.3 Human health for recreation

Specific advice concerning the appropriateness of season and/or flow stratification of sample data used for assessing the non-compulsory Human Health for Recreation primary contact objective.

1.2 Information requirements

The National Policy Statement for Freshwater Management 2014 (NPS-FM) is the Government's national direction for freshwater management. The NPS-FM defines numeric attribute states, which guide the setting of freshwater objectives by numerically describing different levels (states) to which a water body might provide for a given value. The Ministry requires advice on three broad issues associated with using sample statistics to define these numeric attribute states:

1.2.1 Number of samples

The first issue requiring advice is the number of samples and sampling duration needed to characterise the state of a managed environment. The NPS-FM describes some numeric attribute states using sample statistics such as medians, maximums or other percentiles. Most of these descriptions are prefixed by the word "annual", for *example, the nitrate toxicity attribute state is defined by the annual median and annual 95th percentile, and the ammonia toxicity attribute state is defined by the annual median and annual maximum. However, there is no direction concerning the number of samples required to inform estimates of current state.*

There are other attribute states which are not prefixed by the word "annual", and the minimum number of samples and duration is more clearly specified. It is unclear whether these sampling regimes should inform minimum sampling requirements for attributes lacking such specifications. For example, the Planktonic Cyanobacteria attribute state is defined by the 80th percentile of at least 12 samples collected over three years, and the periphyton attribute state is defined as the 92nd percentile of monthly samples collected over at least three years. The Ministry requires guidance on the practical approaches available for councils to assess current state for the range of attributes.

1.2.2 Attribute state switching

A second issue is how the choices made when deciding a sampling regime can affect what the current attribute state of an environment is assessed as. This issue is most likely to arise where uncertainty associated with the estimate of an attribute state is high, and the true state of the environment is near a threshold boundary. High uncertainty could occur if the word "annual" is interpreted to mean that the relevant sample statistic is to be estimated from one year of data, which may mean only 12 or 4 samples are used depending on whether sampling is monthly or quarterly. When the number of samples is limited, there will be high uncertainty associated with the estimated statistic. This imprecision reflects sample variability and potentially inter-annual variation in the drivers of the attribute state. If the site's true status is close to the threshold of interest, the estimated (imprecise) statistic may switch between meeting and not meeting the objective in successive years. This switching of the assessed attribute state is problematic and not helpful to council staff and communities implementing the NPS-FM. It creates issues in establishing the current state (policy CA2) and recognising the importance of 'long-term' trends in monitoring progress towards achieving objectives in plans (policy CB1 of the NPS-FM).

It is noted that for State of Environment (SoE) reporting, regional councils, Land Air Water Aotearoa and the Ministry typically calculate sample statistics, such as the median, using 3-5 years of monthly samples. The rationale for this has been that most SoE monitoring is monthly and therefore 3-5 years of data is considered to represent a reasonable trade-off between site numbers, precision of the estimated statistic (e.g., a median value) and limiting effect of long term trends on the statistic. The Ministry requires advice on how councils can best consider uncertainty in their sampling regime.

1.2.3 Sample statistic for *E. coli*

The third issue relates specifically to the sample statistic for assessing the non-compulsory primary contact objective for Human Health for Recreation. The assessment statistic for primary contact is the 95th percentile. There is a view that stratifying the monitoring data by season and/or flow may be appropriate to restrict the assessment to periods when primary contact recreation occurs. Such stratification would affect the total number of samples in an annual period and could reduce the precision of the statistic. The Ministry requires advice as to the appropriateness and implications of stratifying the monitoring data, and implications for the sampling regime and uncertainty.

1.3 Approach taken in this report

Chapter 2 addresses some fundamental statistical issues and possible means of their resolution and implementation. In that chapter **bold** words signify key concepts. Subsequent chapters address the brief’s topics to be addressed. Footnotes are used liberally, and some more extensive technical material is in the Appendices.

2 The broad statistical issues

2.1 General principles and why they should be understood

Environmental monitoring inevitably impinges on the realm of statistics. So some general understanding of that interaction is desirable, not least because terms may have different meanings in each discipline.[1]

Most importantly, to an environmental professional a **sample** is a volume or mass of material taken from the environment—for example, a container of stream water for subsequent laboratory analysis. But to a statistician, a sample is a collection of results, called **observations**: for example, phosphorus concentrations from a set of physical samples.[2] So to a statistician one **sample** contains **many data** (i.e., observations) and the **sample size** is the number of data in the sample—not the volume of the sampling container. The context usually makes the meaning clear. Similarly, an environmental professional would regard an **error** as just that—a mistake. But to a statistician, **sampling error** is the natural **variability** inherent among data taken from a **population** and is therefore always present and needs to be accounted for (Barnett & O'Hagan 1997).

2.1.1 Statistical population

With existing technology it is often impossible to measure water quality **attributes** in lakes and rivers continuously (temperature, pH and nitrates are exceptions). Therefore we take occasional samples (i.e., make observations). That's because the laboratory effort required for many attributes is substantial (e.g., total nitrogen, dissolved reactive phosphorus), as is the cost of running field parties who perform the sampling. These physical samples are understood to be taken from a population. We make **inferences** about the form of that population using the field and laboratory data we obtain. These inferences concern the shape of the population's statistical **distribution** and its **parameters**, for example the **mean** and **standard deviation** of a **normal** distribution, for which these two parameters reflect the distribution's **central tendency** and **variance** around that.[3]

2.1.2 Statistical "sampling error" and uncertainty

When the population distribution is unknown, the presence of sampling error is inevitable. It is the result of observing a sample instead of the whole population. In precise terms, sampling error is the difference between a **sample statistic** used to estimate a population parameter and the actual *but unknown* value of that parameter.[4]

Physical samples are generally only a tiny fraction of the total water volume about which we wish to make inferences, typically on the order of 1 in a billion. That means that our inferences are necessarily **uncertain**. So when we make inferences about the population from a set of sample data we *need always to be wary of the influence of statistical sampling error*.

More generalised discussion of uncertainty issues is given by Norton et al. (2015).

2.1.3 Accuracy (precision and bias)

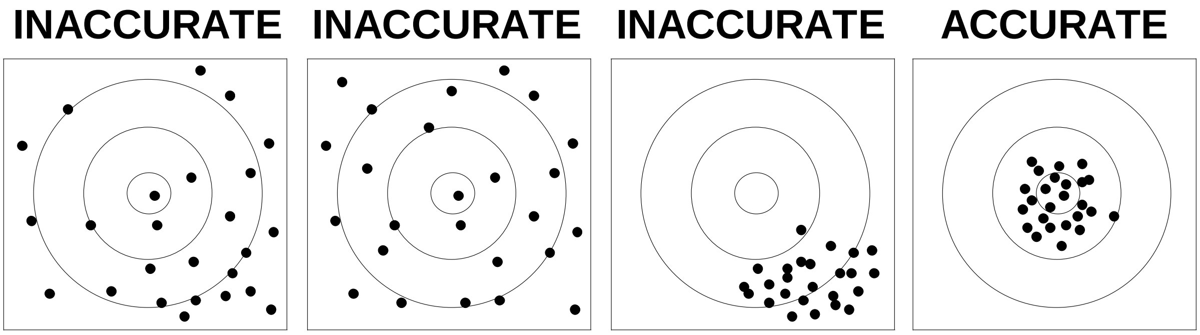
To be accurate, data must be both precise and unbiased, as depicted in Figure 2-1.

[1] The following material is based on McBride (2005).

[2] https://stats.oecd.org/glossary/detail.asp?ID=6132

[3] A statistical distribution can be seen as the form of a frequency histogram as the number of observations becomes huge (see Section 2.1.4).

[4] https://en.wikipedia.org/wiki/Sampling_error



(a) Biased, imprecise (b) Unbiased, imprecise (c) Biased, precise (d) Unbiased, precise

Figure 2-1: Accurate observations are both precise and unbiased. Source: McBride (2005).

Random sampling removes **sampling bias**, but if there are errors in measurement (nothing in this world is perfect), bias can still appear caused by **measurement error**.

Note that many environmental sampling programmes are **systematic**, not random. They are conducted at regular intervals (typically monthly) and often at the same-time-of-day. This is true for data intended to be used for state-of-the-environmental and (especially) for trend analysis. Systematic programmes are chosen for three reasons:

1. To meet the requirements of trend assessment models (most time-series analysis methods demand equally-spaced data)[5].
2. To facilitate efficient sampling by field parties.
3. To reduce known variability.

An example of the last item is fixed-interval sampling of river dissolved oxygen *at the same time-of-day*. River dissolved oxygen often follows a *regular* sinusoidal variation over 24 hours and so known variability is removed when sampling at the same-time-of-day, as depicted in Figure 2-2. Inferences then become less uncertain.

At first glance, systematic sampling compromises accuracy. It certainly does for same-time-of-day sampling of water quality attributes that show regular variations *but only if* inferences were to be made about patterns appearing over the full 24 hours. If inferences were restricted to dissolved oxygen around mid-morning the bias is effectively removed, because the population being sampled has been restrained. Inferences for other times-of-the-day would have to come from other **special investigations**.

The same caveats apply when sampling attributes exhibiting irregular variations. And note that “The researcher must ensure that the chosen sampling interval does not hide a pattern”.[6] For example a factory on a river may discharge cleaning agents only on a Friday. If river sampling is always on a Tuesday that “hidden” pattern will not be sampled—though it could be discovered under strict random sampling.

[5] https://en.wikipedia.org/wiki/Unevenly_spaced_time_series
[6] https://en.wikipedia.org/wiki/Systematic_sampling

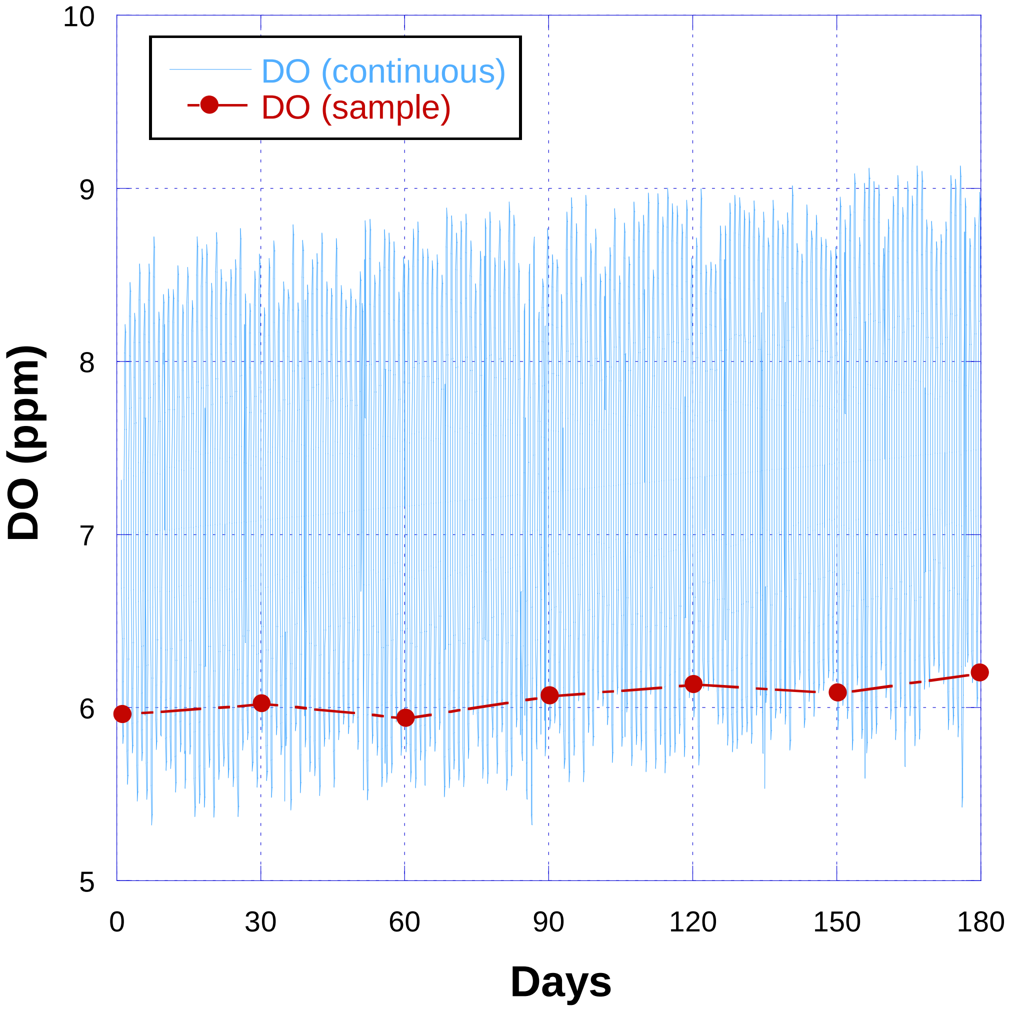


Figure 2-2: Reducing data variance by regular monthly sampling at same-time-of-day (9 a.m.)

2.1.4 Probability distributions

Statistical distributions lie at the heart of much of statistical inference—their shape and parameters must be **estimated**. This requires understanding of the symmetric **normal distribution** (e.g., as may apply to total nitrogen concentrations in lakes), or the right-skewed **lognormal distribution** (especially for *E. coli*).[7] Familiarity with these fundamental features should facilitate ready understanding of many of the following issues.

Populations are expected to be characterised by such statistical distributions. We can use accurate (or at least unbiased) data collected from a population to infer the shape and parameters of that distribution by plotting the observation's frequencies as a histogram, and then imagining what the shape would be were a huge number of accurate samples to be taken. Such a process is depicted in Figure 2-3 (for the normal distribution) and Figure 2-4 (for the lognormal distribution). It gives rise to probability density functions (pdf), as shown by the smooth curves on these figures.

[7] Both these distributions are fully defined given values of the population's median and coefficient of variation (the standard deviation divided by the mean). For low coefficients of variation (e.g., 0.2) the two distributions are quite similar but for high values of that parameter (e.g., 1.0) the two are quite dissimilar, with the lognormal distribution being asymmetric and right-skewed.

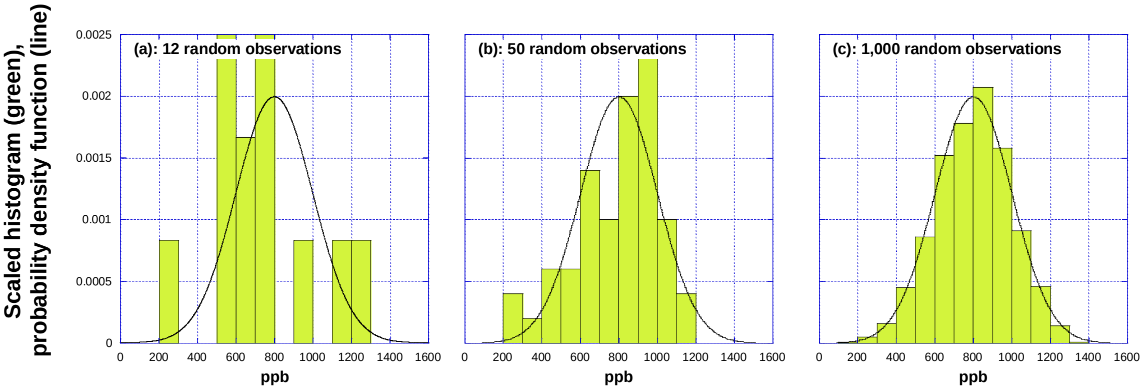


Figure 2-3: From normal histograms to distributions. Random samples drawn from a normal distribution with its two parameters known: (i) median (= mean = mode) = 800 ppb; (ii) coefficient of variation = 25% (so the standard deviation is 200 ppb). The green shaded bins are the relative frequencies divided by their width. For example, the number of observations in the first bin in Figure 2-3(a) is 1 (i.e., only one observation was between 200 and 300 ppb). The total number of data is 12, so the relative frequency is 1/12. The bin width is 100 ppb, and so the scaled histogram height for the first bin on Figure 2-3(a) is 1/(12x100) = 0.000833—because the total area under the bars must be unity. [For Figure 2-3(b) that bin contains 2 data and so its scaled height is 0.0004 and for Figure 2-3(c) that bin contains 5 data and so its scaled height is considerably diminished (0.00005).]

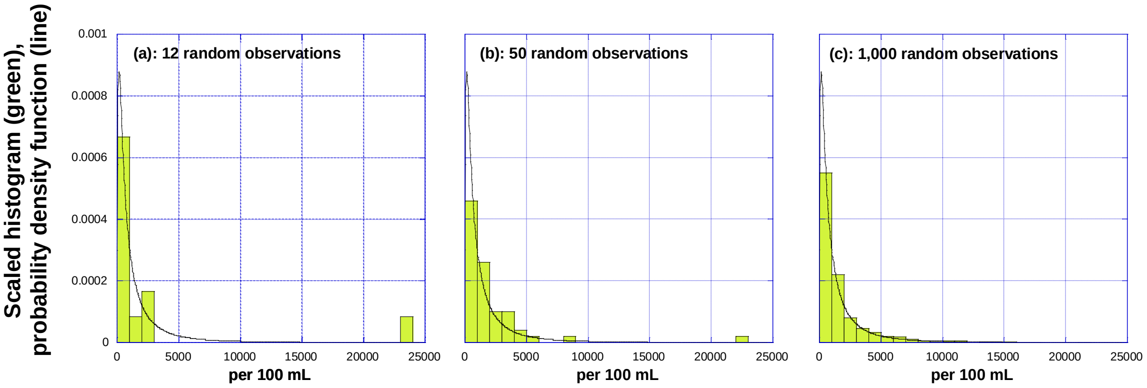


Figure 2-4: From lognormal histograms to distributions. Random samples drawn from a lognormal distribution with its two parameters known: (i) median = 800 per 100 mL; (ii) coefficient of variation = 200% (so standard deviation of natural logarithms of concentrations = 1.27).[8] From these data we calculate mode = 160 (where the pdf is 0.00088), and mean = 1789 per 100 mL.[9] The green shaded bins are the relative frequencies divided by their width. For example, the number of observations in the first bin in Figure 2-4(a) is 8 (i.e., 8 observations are between 0 and 1,000 per 100 mL). The number of data is 12, so the relative frequency is 2/3. The bin width is 1000 per 100 mL, and so the scaled histogram height is 0.000666 [for Figure 2-4(b) that bin contains 23 data and so its scaled height is 0.00046 and for Figure 2-4(c) the first bin contains 548 data and so its scaled height is 0.000548].

[8] Approximate value for many Freshwater Microbiological Research Programme sites' *E. coli* data (McBride et al. 2002, Till et al. 2008).

[9] The mode is calculated from $\exp(\mu_y - \sigma_y^2)$ and the mean is $\exp(\mu_y + ½\sigma_y^2)$, where μ_y is the mean of (natural) logarithms of the data and σ_y is their standard deviation, calculated from $\sqrt{[\ell n(1 + \eta^2)]}$, and η is the coefficient of variation of the raw (non-transformed) data—Gilbert (1987, p. 156); Millard (1998, p. 149).

From Figure 2-3 and Figure 2-4 notice that:

1. Each of these distributions can be characterised by two variables: (i) median and (ii) coefficient of variation (the standard deviation divided by the mean).[10]

2. The total area under each histogram or density curve is one.

3. The height of a density curve is *not* a probability; probabilities are given by *areas under the density curve.* This means that for continuous attributes (such as total nitrogen) the probability that it takes a specific value is virtually zero (because the width of a bin bounded above and below by that specific value is zero)—even though TN must be *a* value.

4. In order to preserve the unit area under the histograms we see a pattern of "unders" = "overs". For example, in Figure 2-3(b) compare the 700–800 ppb bin with the 900–1000 ppb bin.

5. With few data, samples from a strongly skewed lognormal distribution can throw up occasional high values such that the histogram bar can lie considerably above the density curve [Figure 2-4(a)]. However, the more random data we have the closer the histogram mimics the continuous distribution line [compare (a), (b) and (c) panels in these Figures].

6. Normal distributions can intrude into the negative horizontal axis—lognormal distributions can't.[11]

7. For lognormal distributions the median is a much better indicator of **central tendency** compared to the mode (much too low) or the mean (much too high).

Note that these results are obtained using a known distribution with known parameters. In environmental science we seldom have such luxury and so the form of the distribution and its parameters must be inferred from data. With few data such inference is quite uncertain. Even with 50 samples uncertainty remains: examining the histogram alone in Figure 2-3(b) might just as well hint at a left-skewed distribution as it does the truth (i.e., those samples were drawn at random from a normal distribution).

2.1.5 Parametric versus non-parametric methods

As we have seen, some water quality data can have pronounced **skew**, in that occasional very high values are observed whereas most of the time the observations are much smaller (e.g., *E. coli* in rivers, Figure 2-4). In many cases this can be accommodated by using methods that are based on skewed distributions (particularly lognormal, sometimes also the gamma distribution). But in other cases there can be too much uncertainty in the choice of distribution, particularly when there are only small sets of observations. In such cases it can be attractive to use a class of statistical methods that require rather fewer assumptions about distributions. These are the "non-parametric" methods. They are not completely free of the need to assume certain distributions,[12] but nearly so. Their use in multiple-site multiple-attribute trend assessments hugely simplifies the effort—by not requiring an

[10] Usually the lognormal distribution is characterised by the mean and standard deviation of the natural logarithms of the data. These are not as easily grasped as are the median and coefficient of variation of the raw (not transformed) data. So the result from mathematical statistics that these logarithm quantities can be replaced by such well-understood quantities is as remarkable as it is helpful.

[11] They can produce a negative result if the distributions two parameters are augmented by a third "shift" parameter.

[12] For example, the well-used "Wilcoxon Signed Ranks Test" assumes that the distribution of differences between paired samples are symmetric (Conover 1980, p. 281).

examination of each and every case for an appropriate distribution (which, as noted, can be very uncertain anyway).

In general these non-parametric procedures operate on the ranks of data rather than their actual magnitude. So they make inferences about medians whereas parametric methods would make inferences about means.

Nonparametric methods are less **powerful** than parametric methods, *if* the latter's assumptions are met.[13] If that is not the case (e.g., applying normal distribution's results to a lognormal distribution population) the non-parametric method's results should be relied upon. These methods are therefore often used in water-related environmental statistical methods, especially for trend analyses, because parametric assumptions may be inappropriate.

2.1.6 Confidence intervals

C**onfidence intervals** are ranges within which a parameter (e.g., median, a percentile) may lie most of the time, under repetitive sampling. Statistical theory enables their calculation, parametric or non-parametric, which is generally straightforward. They may be **two-sided** (a finite range) or **one-sided** (greater than a stated value, or less than that value). Two-sided confidence intervals tend to shrink as the sample size is increased.[14] Indeed for an infinite sample size these intervals have zero width, because the population parameter is then known exactly.

See Appendix A for a more detailed discussion of their interpretation, in the light of the "repetitive sampling" requirement in the previous paragraph.

2.1.7 Percentiles

Percentiles of contaminant concentrations are increasingly used in water management. Merely stating a maximum-not-to-be-exceeded is insufficient to characterise the environment in which communities want aquatic life to be safeguarded. Aquatic organisms "see" time-histories of the concentration of contaminants and we need to characterise usual and unusual concentrations in that time-history, using percentiles.

In particular, a percentile indicates the value below which a given percentage of data fall. Those data can be actual observations (sampling result) *or* the **random variable** of a distribution. For example, the sample 80th percentile (denoted herein as 80%ile) is the value below which 80 percent of observations may be found, but this will always be different from the 80%ile of the population from which samples have been drawn! That outcome is the result of uncertainty—statistical sampling error. It is therefore, important to always make clear whether a percentile refers to observations or to populations. This will become more clear when we consider **hypothesis testing** (section 2.1.10) and **burden-of-proof** (section 2.1.11).

Note that there is no one correct way to calculate percentiles. Software help files often fail to alert users to this fact. Percentile calculation methods generally involve an interpolation between adjacent ranked data. For example, if we want to calculate a 95%ile from 22 data, ranked from lowest to highest, which datum or data should we use? Should that be the highest (22nd), the next highest (21st) or some combination thereof? What if we had 12 samples and wish to estimate the 80%ile. Should that be the 9th or 10th highest or something in between?

[13] Less power implies that statistical assessment of data will be more uncertain.

[14] The more samples we have the more confident we can be about our estimates (e.g., of a distribution's mean).

Three methods are often used to calculate sample percentiles: Excel, Weibull and Hazen (Ellis 1989, McBride 2005, chapter 8). For 22 data, these calculate the 95%ile as the datum with rank 20.95 (Excel), 21.4 (Hazen), 21.85 (Weibull), so interpolation between adjacent ranked data is required. One can see that if the 21st and highest data differ substantially, markedly different 95%ile values may be calculated. Examples of how to do these calculations are given in a footnote.[15]

For any valid sample size the Weibull result always exceeds the Hazen result which always exceeds Excel result. International practice for microbiological water quality (WHO 2003), and the New Zealand primary contact guidelines (MfE/MoH 2003), use the Hazen method, as a half-way house between the Excel and Weibull methods.[16] Note that the Hazen method requires a minimum of 10 data to be able to calculate a 95%ile; Weibull needs 19 and the Excel method requires only one![17]

2.1.8 Tolerance intervals

A tolerance interval limit is effectively a percentile inflated or deflated a little to take account of statistical sampling error.[18] They can be one-sided or two-sided, but the one-sided version is particularly appropriate for the NOF. Using these intervals, instead of using percentiles directly, allows for the influence of statistical sampling error. In doing so, deciding on the appropriate burden of proof will be necessary, as discussed in section 2.1.11. Their use in the NOF for *E. coli* seems unwarranted as the burden-of-proof has already been decided (see section 5.1.2), but they could be considered for other attributes.

Appendix D of McBride (2014) shows that for a small sample size (n = 12) the 95% upper one-sided tolerance interval (for a coverage of 95%) is greater than the Hazen 95%ile, as may be expected. However for a larger sample size (n = 60) these two quantities are very similar in value, as expected. Further technical details are given in Appendix B.

2.1.9 Misclassification error risk

As an extreme case of the effects of statistical sampling error consider lake TP, where the A/B attribute band threshold is an annual median of 10 ppb.[19] Let's say that year after year the lake's concentration history just qualifies it for Attribute A, in that its true (but unknown[20]) annual median concentration was always in fact 9.99 parts per billion. If we take 13 samples each year from that lake (i.e., one every four weeks), what may we find? Standard theory says that about half of those years the sample median (the seventh highest value) will be greater than 10 and so the lake would be misclassified for about half of the years assessed. This is the effect of the statistical sampling error, reflected as "unders and overs", denoting the effect of uncertainty when estimating the lake's annual TP concentration. That value (9.99 ppb) will seldom be attained exactly in the results of a sampling programme—precisely because of statistical sampling error. If the "annual median" is interpreted as the median of sample values the response of many can be to translate "annual median" in the

[15] Let p = percentile fraction (e.g., p = 0.95) and n the sample size (e.g., n = 22). Using the formula embedded in Excel, the rank of the 95%ile is $r = 1 + p(n–1) = 20.95$. Using linear interpolation between adjacent ranked data this value is the weighted average of the 20th and 21st ranked data, with weight of 0.05 on the 20th ranked datum and weight 0.95 on the 21st ranked datum. For Hazen the ranking formula is $r = ½ + pn = 21.4$, while the Weibull formula is $r = p(n+1) = 21.85$.

[16] A Hazen percentile estimator can be found at http://www.mfe.govt.nz/publications/fresh-water/bathewatch-user-guide/hazen-percentile-calculator. It may be possible to simplify this estimator in the form of a UDF in Excel (User Defined Function).

[17] The minimum sample size formulae are as follows. Weibull: $p/(1–p)$. Hazen: $1/(2p)$. Excel requires only one datum (an extremely undesirable property; any number is its own percentile of any order!)

[18] Inflation is appropriate if you want to be very sure that the true percentile is below the limit, in which case we would use an upper one-sided tolerance limit on the 95%ile.

[19] Refer to the Attribute Table on Page 26 in NZGovernment (2014).

[20] Indeed this concentration is unknowable.

percentile statement into medians *of samples*. This therefore accepts (often unwittingly) a misclassification risk as high as 50%.

2.1.10 Hypothesis testing

Hypothesis tests aim to see whether we can give credence to a stated hypothesis, or its alternative, given new data. Many tests in current use address what some have called the “nil hypothesis” (Cohen 1994) because they test the notion that there is no (‘nil’) difference whatsoever between a parameter and some stated value. As we have already seen (section 2.1.4) there is virtually no chance that such a hypothesis could be true. Not surprisingly, there are therefore many problems in this approach (McBride et al. 2014).[21] Fortunately, these generally do not arise in the context of the National Objectives Framework—because the issues to be faced are one-sided—has a threshold been exceeded or not?

One-sided tests nicely portray the burden-of-proof issue (considered in more depth in the next section). Let’s take lake TP as an example again and interpret its A/B threshold as a median of *time* (over a year), not necessarily of *samples*. If we take a *precautionary approach* we would test the hypothesis that the true annual median phosphorus concentration is *greater than* 10 ppb, i.e., it is in attribute states B, C or D. We would only reject that hypothesis if data are sufficiently convincing to do so.[22] And if we did we would infer attribute state A. On the other hand, if we take a *permissive approach* we would test the hypothesis that the true annual median phosphorus concentration is *less than* 10 ppb, i.e., it is in attribute state A. We would only reject that hypothesis if data are sufficiently convincing to do so.[23] And if we did, we would infer attribute state B, C or D. So the form of the test used has everything to do with the burden-of-proof.

One-sided approaches can also be used in the context of assessing trend direction, a notion very consistent with the idea of “progress assessment”. McBride et al. (2015) and Larned et al. (2015, 2016) present and utilise a “two one-sided” procedure that first seeks to identify the trend direction. It only goes on to consider its *environmental* significance (rather than its *statistical* significance) if the trend direction can be confidently inferred. If that can’t be done one infers that there are insufficient data or that trends have reversed during the period of record. This trend direction-detection procedure uses the Greek symbol alpha (α) to denote the maximum permissible misclassification error risk,[24] whereas the *P*-value-based hypothesis test procedure uses α as the “significance level”, the maximum “Type I error rate”.[25] See McBride et al. (2015) for a discussion of all the subtleties involved. (It should be noted that this procedure has yet to be fully evaluated by independent statisticians.)[26]

2.1.11 Burden-of-proof

This matter is discussed at some length by McBride (2014). The issue concerns how we account for sampling error when assessing the true Attribute State. As discussed in section 2.1.10, if the true

[21] These issues have been well-discussed in the statistical literature (http://warnercnr.colostate.edu/~anderson/thompson1.html), but they seldom find their way into applied science texts.

[22] The sample median would have to be somewhat less than 10 ppb for that to occur, wherein “somewhat” has to do with sample size—the more data we have the smaller that “somewhat” would be.

[23] The sample median would have to be somewhat *greater* than 10 ppb for that to occur. Again, the more data we have the smaller that “somewhat” would be.

[24] Falsely inferring a negative trend direction, and vice versa.

[25] A type I error occurs when the analyst rejects a true hypothesis.

[26] It should also be noted that the presentation of this trend direction assessment procedure was presented in McBride et al. (2015) (Appendix A of the Larned et al. 2015 report) as a form of hypothesis test. On reflection, it should not be so-regarded. Even though its mechanics are somewhat similar to test procedures (especially the TOST, Two One-Sided Test for interval hypotheses, McBride 2005, sec. 5.3.2), it is an *assessment procedure*, not a *test*.

state was close to a NOF threshold, misclassification error risks can rise to as much as 50%. So when we make an assessment, we have three options to account for the sampling error:

1. Ignore it, in which case we take an *even-handed approach* ("face-value") to misclassification error risks, so they are uncontrolled. In this case percentiles of samples are taken directly as percentiles of time.

2. Take a *permissive approach*, by assuming that a NOF threshold has *not* been exceeded and only abandoning that assumption if data become sufficiently convincing (controlling the risk of inferring a lesser state than is actually the case). This approach can also be given the labels: "proof of hazard", "slipping through the net", "letting the guilty go free" or "benefit of doubt". In this case a sample percentile is lower than the even-handed result.

3. Take a *precautionary approach*, by assuming that a NOF threshold *has* been exceeded and only abandoning that assumption if data become sufficiently convincing (controlling the risk of inferring a higher state than is actually the case). This approach can also be given the labels: "proof of safety" or "fail-safe". In this case a sample percentile is higher than the even-handed result.

It is not clear if development of the current NOF tables gave any attention to this issue,[27] so most readers will assume the even-handed approach and take data at face-value, interpreting percentiles as *percentiles of samples*, and ignoring or accepting uncontrolled misclassification error risks. But if the NOF percentiles are interpreted as *percentiles of time* then we need to use data to make inferences about those values, using one of the three burdens-of-proof listed above. I imagine few would advocate a permissive approach but some may favour more emphasis on the precautionary approach.

2.2 Design

2.2.1 Statement of intent—target population

This is essentially covered in section 2.1.3. So, for example, if lake samples are always taken in deep water, that should be stated in appropriate protocols, to alert others to the fact that inferences made from that site may not validly translate to samples taken from shallow waters at the lake margins. Or, as earlier, since many sampling runs result in samples being collected at the same-time-of-day, that too needs to be stated, especially for variables that can vary substantially over the day-night cycle (e.g., DO, pH, water temperature).

2.2.2 How many samples?

Various formulae can be adduced for this question. An earlier report (McBride 2014) suggested an alternative approach:

> *Estimates of medians and 95%iles become more precise as the number of samples is increased, assuming that the sampling programme is bias-free. One can calculate the number of samples needed to meet nominated particular confidence limits, and so select a particular "sample size". While this approach has the attraction of apparent objectivity, it is also rather arbitrary—depending on the particular chosen confidence level.*

[27] Most NOF tables do not make a distinction between sample percentiles and population percentiles. The only one that does is river periphyton, where sample percentiles are mandated.

Another (semi-quantitative) approach is to examine the overall properties of the confidence limit curve, as a function of the number of samples. Such curves are given in the figures below.[28] *They indicate that somewhere in the region of 20–40 samples, one reaches an area of rapidly diminishing returns. Whilst being semi-quantitative (appealing to the general shape of the confidence limit curves) this seems a more informed manner in which to decide on an appropriate sample size.*

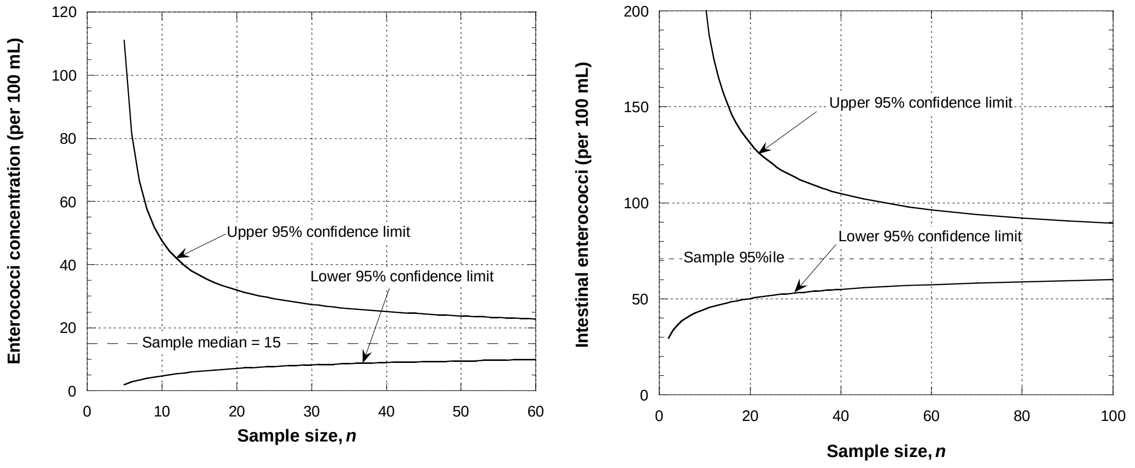


Figure 2-5: Lognormal confidence limits for the median and 95%ile. Source: McBride (2005), Figures 3.1 and 3.4. (Without loss of generality, these figures were prepared for enterococci concentrations.)

Such a presentation implies that once 40–60 samples have been obtained, precision may be satisfactory. WHO (2003, page 83) suggests a minimum of 60 samples for microbial water quality assessment. Such numbers may be taken as a useful guide to selecting a desirable sample size.

Another consideration when deciding on an appropriate sample size is "state switching frequency", where once again we have to consider the effects of statistical sampling error. This topic is analysed in section 3.1.

2.2.3 Sampling and analysis protocol

Good documentation of sampling sites is vital. For example, when interpreting river data, readers should be alerted to the fact that sampling is from the midpoint of a bridge spanning the river, and is taken at the surface, or at some stated depth. Details of the field and laboratory records should also be listed. Many aspects of what could and should be included in such a protocol have been covered elsewhere (Ward et al. 1990, Davies-Colley et al. 2012).

2.2.4 Should we use sample percentiles, tolerance limits or a look-up table?

If an even-handed approach is taken when assessing Attribute States, sample percentiles should be used directly, ideally using the Hazen formula. This is the simplest approach.

[28] Without loss of generality, these graphs are for enterococci concentrations, distributed as lognormal. More-or-less the same pattern can be expected for other water quality variables.

If a precautionary approach is taken, necessarily interpreting percentiles as percentiles of time (not percentiles of samples), one-sided tolerance limits could be used,[29] but with one exception: Primary contact recreation is already based on a precautionary approach, as discussed later (section 5.1.2).

However this tolerance limit approach requires assumptions about the distribution of the Attribute (normal, lognormal, gamma,…) and it can be simpler, and possibly more powerful, to take a non-parametric approach, such as is taken in assessing the state of drinking-water supplies.[30] This uses "look-up" tables, presenting the permissible number of exceedances of a threshold in a given number of samples, keeping the precautionary misclassification error risk below 5%. The theory behind this approach is reported in McBride & Ellis (1991), and discussed in McBride (2014). Results for 95%iles are shown on Table 2-1.

2.2.5 Look-up tables for the precautionary approach

The following three tables display lookup tables for the precautionary approach.

Table 2-1: Look-up table for allowable exceedances in a precautionary approach to 95%ile assessment.

e	*n*
0	38–76
1	77–108
2	109–138

Note: '*e*' is the maximum permissible number of exceedances of a 95 percentile threshold for the stated range of samples '*n*', with maximum misclassification error risk of 5%. Calculations have been made using the theory stated in McBride and Ellis (2001), using 'Jeffreys' prior'. (See also McBride 2005, Table 8.1.) These numbers are little changed when assuming different prior distributions or more elaborate parametric models (McBride 2003). Note that if there are less than 38 samples it is not possible to keep misclassification error risks below 5% when assessing a 95%ile standard.

In this Table we see the effect of the precautionary approach. There can be no exceedances if there are less than 77 samples, and so the Numeric Attribute State is effectively a maximum. If the true proportion of time that the 95%ile threshold was exceeded was indeed 5%, at most 1.8% of samples could exceed this threshold (i.e., 2/109). That's because a precautionary approach has been taken. Were a permissive approach to be taken, more than 5% of samples could exceed the threshold (e.g., with 100 samples it is 9%—see Table 2-4.

Table 2-2 displays the look-up table for 80%iles (used in the NOF table for cyanobacteria) and Table 2-3 gives the table for medians (used in a number of NOF tables).

[29] And as noted in Appendix B, one-sided tolerance limits on percentiles are identical to confidence limits on percentiles. (This is not the case for two-sided intervals.)

[30] See section 6.2.2 at http://www.health.govt.nz/system/files/documents/publications/guidelines-drinking-water-quality-management-for-new-zealand-2015-oct15.pdf.

Table 2-2: Look-up table for allowable exceedances in a precautionary approach to 80%ile assessment

e	*n*	*e*	*n*	*e*	*n*
0	9–17	3	33–39	6	53–59
1	18–25	4	40–46	7	60–65
2	26–32	5	47–52	8	66–71

Note: '*e*' is the maximum permissible number of exceedances of a 80%ile threshold for the stated range of samples '*n*' with maximum misclassification error risk of 5%. Calculations have been made using the theory stated in McBride and Ellis (2001). using 'Jeffreys' prior'. Note that with fewer than 9 samples it is not possible to achieve 95% confidence that the 80%ile has been met, even if there were no exceedances. These numbers have not been published previously; they were calculated using the author's Fortran program "**Concom**". Note that if there are less than 9 samples it is not possible to keep misclassification error risks below 5% when assessing a 80%ile standard.

Table 2-3: Look-up table for allowable exceedances in a precautionary approach to 50%ile assessment

e	*n*	*e*	*n*	*e*	*n*
0	3–5	8	25–26	16	43–44
1	6–8	9	27–28	17	45–47
2	9–11	10	29–31	18	48–49
3	12–14	11	32–33	19	50–51
4	15–16	12	34–35	20	52–54
5	17–19	13	36–38	21	55–56
6	20–21	14	39–40	22	57–58
7	22–24	15	41–42	23	59–60

Note: '*e*' is the maximum permissible number of exceedances of a median threshold for the stated range of samples '*n*' with maximum misclassification error risk of 5%. Calculations have been made using the theory stated in McBride and Ellis (2001), using 'Jeffreys' prior'. These numbers have not been published previously; they were calculated using the author's Fortran program "**Concom**". Note that if there are less than 3 samples it is not possible to keep misclassification error risks below 5% when assessing a 50%ile standard.

2.2.6 Look-up tables for the permissive approach

The following three tables display lookup tables for the permissive approach.

Table 2-4: Look-up table for allowable exceedances in a permissive approach to 95%ile assessment.

e	*n*	*e*	*n*	*e*	*n*
0	1–3	3	23–34	6	61–74
1	4–11	4	35–46	7	75–88
2	12–22	5	47–60	8	89–102

Note: '*e*' is the maximum permissible number of exceedances of a 95 percentile threshold for the stated range of samples '*n*', with maximum misclassification error risk of 5%. Calculations have been made using the theory stated in McBride and Ellis (2001), using 'Jeffreys' prior'. (See also McBride 2005, Table 8.1, but note that the above table contains slight differences.) These numbers are little changed when assuming different prior distributions or more elaborate parametric models (McBride 2003). They were calculated using the author's Fortran program "**Concom**".

In this Table we see the effect of the permissive approach. If only one sample is to hand, exceeding the threshold, it may not be inferred that the threshold has been exceeded for 95% of the time!

Table 2-5 displays the permissive look-up table for 80%iles (used in the NOF table for cyanobacteria) and Table 2-6 gives the table for medians (used in a number of NOF tables).

Table 2-5: Look-up table for allowable exceedances in a permissive approach to 80%ile assessment

e	*n*	*e*	*n*	*e*	*n*
0	1	9	28–30	18	64–67
1	2–3	10	31–34	19	68–71
2	4–6	11	35–38	20	72–75
3	7–9	12	39–42	21	76–79
4	10–12	13	43–46	22	80–84
5	13–16	14	47–50	23	85–88
6	17–19	15	51–54	24	89–92
7	20–23	16	55–58	25	93–96
8	24–27	17	59–63	26	97–101

Note: '*e*' is the maximum permissible number of exceedances of a 80%ile threshold for the stated range of samples '*n*' with maximum misclassification error risk of 5%. Calculations have been made using the theory stated in McBride and Ellis (2001). using 'Jeffreys' prior'. These numbers have not been published previously; they were calculated using the author's Fortran program "**Concom**".

Table 2-6: Look-up table for allowable exceedances in a permissive approach to 50%ile assessment

e	*n*	*e*	*n*	*e*	*n*
0	1	10	14–15	20	31–32
1	1	11	16–17	21	33–34
2	2–3	12	18	22	35–36
3	4	13	19–20	23	37
4	5–6	14	21–22	24	38–39
5	7	15	23	25	40–41
6	8–9	16	24–25	26	42–43
7	10	17	26–27	27	44–45
8	11–12	18	28–29	28	46
9	13	19	30	29	47–48

Note: '*e*' is the maximum permissible number of exceedances of a median threshold for the stated range of samples '*n*' with maximum misclassification error risk of 5%. Calculations have been made using the theory stated in McBride and Ellis (2001), using 'Jeffreys' prior'. These numbers have not been published previously; they were calculated using the author's Fortran program "**Concom**".

These tables dramatically demonstrate why one eminent statistician has titled his paper: "Why proof of safety is much more difficult than proof of hazard" (Bross 1985). Most spectacularly, if 60 samples were to be obtained, even if only one of them exceeded a threshold we cannot be confident (at the 95% level) that the true 95%ile was below that threshold. For the median assessment once we get 24 or more exceedances out of 60 samples (≥40%) we cannot be confident that the true median was below the threshold. The tables can be recalculated for different misclassification error risks.

2.3 Implementation

2.3.1 Flexibility in the NOF tables?

Most of the current NOF tables do not explicitly require percentiles *of samples*. The exception is the Table for periphyton in rivers: "Exceeded in no more than 8% of samples" (Default Class) and "Exceeded in no more than 17% of samples" (Productive Class). Less straightforwardly, the cyanobacteria table can also be interpreted as requirements on percentiles of samples (see the footnote to that Table). So there is flexibility in the interpretation of most of the NOF percentiles.

2.3.2 State-switching

This is discussed at some length in section 3.1, where it is concluded that an assessment period of one year is insufficient. It seems preferable to use a five-year period with rolling annual assessments.

2.3.3 Implications for testing for progress assessment

As an example we again consider trends in total phosphorus concentration (TP) in a lake. Traditional trend hypothesis tests can propose that there has been no change in the lake's median TP over time.

That is, no change whatsoever. Samples are taken from that lake and used in a prescribed calculation procedure to see if the hypothesis should be rejected. If it is, a “statistically significant” result is announced, as a consequence of the test’s calculated *P*-value being less than 0.05. If the hypothesis is not rejected one *cannot* go on to accept the hypothesis as being probably true, or that the situation is ‘stable’. All one can say as a matter of inference is that the hypothesis is ‘not rejected’. A major problem with this procedure is that it tests a hypothesis that is always false! There is always change, even if only “small”. That is why a recent New Zealand analysis (McBride et al. 2015) has developed a new assessment procedure in place of a traditional hypothesis test, considering the trend *direction*.[31] Three outcomes are possible: (i) Confidence that the trend is upward; (ii) Confidence that the trend is downward; (iii) There are not enough data to infer the trend’s direction, or the trend may have reversed during the period of record. Many communities, including Maori, want to know if we can confidently infer the trend direction, which is much more informative information than a poorly-understood and not-very-relevant notion of statistical significance.

The direction-of-trend assessment procedure in McBride et al. (2015) is relatively straightforward to implement.[32]

[31] The fundamental ideas for this approach are presented in McBride et al. (2014) and follow from an insightful paper by Jones & Tukey (2000).

[32] It consists of calculating 100(1–2α)% confidence limits about a trend slope estimate. If both limits are positive an upward trend has been confidently inferred (with (1–α)% confidence—*not* (1–2α)% confidence), and vice versa. If the upper limit is positive but the lower is negative then there are not enough data to confidently infer the trend direction or the trend has reversed.

3 Interpreting current NOF statistics

In order to examine the effects of sample imprecision, it is appropriate to first consider how "State switching" may arise.

3.1 State switching and the assessment period

A major purpose of sampling a water body, making measurements of concentrations, and using the date in an assessment, is to communicate its state. Where a state assessment indicates changes over time it is desirable that this be a reflection of real changes in the water body's status and not due to inadequate sample size. This has implications for assessment periods and assessment frequency.

Consider the true course of total nitrogen concentration at a polymictic lake site over a period of a year, and imagine that over that time its true median concentration was 400 mg/m^3. This would qualify the lake for Attribute State B (because the annual median is between 300 and 500 mg/m^3). Of course we never know that true (population) value; all we may have is a set of 12 monthly samples and so the effect of 'statistical sampling error' must be considered. Therefore the median of the twelve samples will not necessarily lie between 300 and 500 mg/m^3. It is entirely feasible for the sample median to be below 300 mg/m^3, in which case it would qualify as Attribute State A. It is also entirely feasible for the sample median to be above 500 mg/m^3 (but less than 800 mg/m^3), in which case it would qualify as Attribute State C. It is even possible (but highly unlikely) for that median to be above 800 mg/m^3, in which case it would qualify as Attribute State D.

It follows that a series of annual assessments of lake medians could involve some "switching" when the true median was always, in fact, reflecting Attribute State B. For example, consider the case where the true (but unknown) annual median was around 350 mg/m^3 for each of ten years. Over a ten-year period we could obtain the sequence B-B-A-B-C-B-B-B-B-B which presents four switches.

Now imagine that the true median total nitrogen concentration over the ten years was much closer to the A/B boundary, say 310 mg/m^3. Statistical sampling error means that switching between States A and B is rather more likely. For example, we could have B-B-A-B-A-A-B-A-B-A, which presents seven switches.

Remedies for this situation include increasing the sampling frequency, lengthening the assessment period (e.g., to three years, or even to five years[33]) and performing rolling assessments. Note that even though rolling assessments can only be performed if the *assessment period* is greater than one year, the *assessment frequency* can still be annual. For example, if the assessment period is five years then the annual assessment would be based on the preceding five years' data. So as a new year's data is added to the dataset, the oldest year's data (now six years old) is deleted and a reassessment is made, as depicted in Table 3-1.

[33] In "grading" recreational water sites the MfE/MoH (2003) water quality guidelines use a period of five years with weekly sampling during the bathing season.

Table 3-1: Data selection regime for annual assessments using five years' data. The median is best calculated for all data in the five year period, rather than taking the median of each year's median.[34]

Year (since commencement of an assessment regime)								
1	2	3	4	5	6	7	8	9
√	√	√	√	√				
x	√	√	√	√	√			
x	x	√	√	√	√	√		
x	x	x	√	√	√	√	√	
x	x	x	x	√	√	√	√	√

"√" denotes data to be used in the assessment; "x" denotes data that are available but are not used; blank cells denote "no data".

It is also possible to use 'adjacent' assessment periods. For example, if an assessment period of three years is selected, assessment would be done at:

- year 3 (using data from years 1–3)
- year 6 (using data from years 4–6), etc.

The effect of these potential remedies can be addressed using numerical experiments in Monte Carlo modelling, in which we take 1000 repetitive samples from a known distribution for each of:

- weekly and monthly sampling over a 15 year period
- assessment periods of 1, 3 and 5 years
- rolling annual assessments (necessarily for 3 and 5 year assessment periods only)
- adjacent assessments (for each of the 15 years).

These analyses have been carried out in an Excel-VBA programme developed specifically for this task.[35] Results are presented below.

3.1.1 Normal (symmetric) distribution results

Table 3-2 presents results for a normal distribution using lake TN as an example.[36] As expected from the discussion above, the closer the true median is to the A/B or B/C boundaries, the higher the switching frequency. In the middle of the Attribute State (400 mg/m^3) there is practically no switching. We also see that, apart from this middle-of-the-range result, annual assessments have the highest frequencies as is also the case for monthly sampling (cf. weekly sampling). Rolling assessments confer lower switching frequencies (cf. adjacent frequencies).

[34] The median of medians is not (quite) the same result as the median of all data.

[35] Latest version: file **State Switching in the NOF, 4 July 2016.xlsm**, developed by the author.

[36] Examination of various lake datasets collected by Dr Noel Burns (Burns & Rutherford 1998) reveals that some lakes TN concentration may be adequately described by a normal distribution (e.g., Lake Rotorua), with CoV rather less than 50% (e.g., ~30% for Lake Rotorua).

Table 3-2: Mean switching frequencies (%) for a lake whose TN concentration confers Attribute State B (annual median between 300 and 500 mg/m^3), where increasing colour intensity denotes proximity to the A/B and B/C Attribute State boundaries.

Case[b]	True median polymictic lake TN (mg/m^3)[a]								
	310	320	340	360	400	440	460	480	490
NW1A	39	20	2	0	0	1	9	33	46
NW3A	23	4	0	0	0	0	0	14	38
NW3R	14	3	0	0	0	0	0	9	22
NW5A	15	1	0	0	0	0	0	8	33
NW5R	7	1	0	0	0	0	0	4	15
NM1A	47	39	22	10	3	17	32	46	49
NM3A	43	26	6	0	0	3	14	37	48
NM3R	25	17	5	1	0	3	11	22	27
NM5A	37	17	2	0	0	0	6	34	46
NM5R	16	9	1	0	0	0	4	14	19

[a] Lake TN concentration is assumed to follow a normal distribution with 30% coefficient of variation.

[b] Case codes: "N" denotes Normal distribution; "W" or "M" denotes Weekly or Monthly sampling; "1", "3" or "5" denotes assessment period; "A" or "R" denotes Adjacent or Rolling assessments

Given that weekly sampling may be impractical, an optimum regime for normally distributed data appears as annual assessments using monthly sampling in a rolling five-year assessment window.

3.1.2 Lognormal (asymmetric) distribution results

Table 3-3 presents results for a lognormal distribution using river *E. coli* concentration as an example.[37] This time, because of the presence of asymmetry (strong right skew), the patterns are a little more complex.

Annual assessment period is again the worst performer. For weekly sampling the switching frequencies are again minimal in the middle of the Attribute State range, but this is nowhere as evident for monthly sampling. Indeed for annual assessment period with monthly samples the switching frequencies slightly increase from left to right, as more "C" State medians are encountered. Rolling assessments again produce lower switching frequencies (cf. adjacent assessments).

[37] The two-parameter lognormal distribution (Gilbert 1987) has the remarkable property that its coefficient of variation (η) is a function *only* of the standard deviation of the natural logarithms of the concentration data (σ_y, where the *y* subscript denotes logged data). It is independent of the mean (or median). So a 150% coefficient of variation (i.e., η = 1.5) gives rise to $\sigma_y = \surd\{\ell n[(1+ \eta^2)]\} = 1.09$. This is rather lower (and hence presents a less skewed distribution) than the observed value for the all sites in the 1998–2000 FMRP data (i.e., σ_y =1.87, see footnote 8). Nevertheless, some FRMP sites displayed lower σ_y values, as do wider datasets (e.g., for the Wairua, Mangakahia and Hoteo rivers, https://data.mfe.govt.nz/table/2533-river-water-quality-raw-data-by-site-2009-2013/). Furthermore, in recent years higher river concentrations of *E. coli* can be expected to have been reduced in magnitude and frequency, with implementation of the 'Clean Streams Accord'. Hence the adoption of η = 1.5 as a typical coefficient of variation.

Table 3-3: Mean switching frequencies (%) for a river whose *E. coli* concentration confers Attribute State B (annual median between 260 and 540 mg/m^3) where increasing colour intensity denotes proximity to the A/B and B/C Attribute State boundaries. [a]

Case[b]	True median river *E. coli* (#/100 mL)[a]								
	274	288	316	344	400	456	484	512	526
LW1A	48	42	26	14	12	30	41	48	50
LW3A	43	29	7	1	1	10	26	44	48
LW3R	24	17	5	1	0	7	16	23	26
LW5A	40	19	2	0	0	4	17	40	46
LW5R	17	10	1	0	0	2	9	17	19
LM1A	52	56	60	62	63	65	65	66	66
LM3A	45	49	52	46	42	49	52	53	53
LM3R	27	29	31	31	29	32	33	33	33
LM5A	41	48	47	35	26	44	50	50	49
LM5R	17	20	20	17	13	19	20	21	21

[a] River *E. coli* concentration is assumed to follow a lognormal distribution with 150% coefficient of variation

[b] Case codes: "L" denotes Lognormal distribution; "W" or "M" denotes Weekly or Monthly sampling; "1", "3" or "5" denotes assessment period; "A" or "R" denotes Adjacent or Rolling assessments

Again, given that weekly sampling may be impractical, an optimum regime for normally distributed data appears as annual assessments using monthly sampling in a rolling five-year assessment window.

3.2 Implications

Annual assessment period and frequency carries substantial risks of false state-switching. Adopting a rolling regime performing annual assessments on the previous five years of data seems desirable, reducing these risks. It is also in harmony with the NOF's periphyton and cyanobacteria Tables, which call for at least three years of data.

Note however that for rare occurrences (e.g., when a lake hypolimnion goes anoxic, releasing a high spike of phosphorus and ammonia from the sediments, or when there has been a large accidental overflow from a WWTP causing a high *E. coli* spike), adjacent assessments *may* be preferred over rolling assessments. This may be particularly appropriate for high percentiles of any 'spiky' attribute. For example, say that this rare event occurred just once in five years. For rolling assessments (over five years) that spike 'penalises' the results for the other four subsequent years—because such extremes did not then occur. In such a case it may be better to accept a higher risk of state-switching (by choosing adjacent multi-year assessments) in order to avoid the penalty, although this would mean that assessments could not be carried out annually. Or, one could use multi-year rolling assessments for medians and single-year assessments for 95%iles and maxima.

3.3 Include an "Unknown" category?

An Australian reviewer (Dr Rob Goudey) has suggested as follows:

> *The "state switching" problem results from an insistence on routine two-alternative outcomes hypothesis testing. It shows what can happen when we insist on making clear-cut yes/no decisions based on very small samples.*
>
> *If the two-one-sided test (TOST) approach described for trend testing (p15 of document) could be used for testing percentiles, this would result in three possible outcomes (i.e., A, U, B where U = act as if undecided so far). Sequences of years in which water quality is close to the threshold of A and B, rather than exhibiting switching behaviour, might then consist of series of "U". This would make more explicit the indecision in assessing water quality against a percentile limit using small sample sizes. The need to pool several adjacent years of observations before a clear inference is possible might become more obvious.*
>
> *Adopting a three–alternatives decision approach would also place a burden of proof on the monitoring program itself, i.e., the sampling frequency needs to be adequate to allow a decision and to minimise the probability of a directional (Type III) error. If assessed using a confidence interval approach, then the width of the confidence interval can provide a measure of quality of the monitoring program to allow decisions.*

This is a topic worthy of consideration. Its implementation would be as outlined by McBride et al. (2014).

4 Consistent approach to future attributes

To some extent the formulation of NOF criteria will be dependent on peculiarities of the different attributes. Uniform consistency of approach is therefore not to be expected.

The issues that do need addressing when aiming for at least some consistency are:

1. Explicit account of whether percentile metrics in the NOF Tables should be considered as percentiles of sample results or percentiles of time over the assessment period.

2. If the latter, the appropriate burden-of-proof and misclassification error risks should be identified, including examination of practicable sampling regimes (in terms of sample size and sampling frequency).

3. Adoption of rolling multi-year assessment periods and an annual assessment frequency—to substantially reduce the misclassification errors that will arise when using single-year assessment periods. Note that so long as rolling assessments are employed, adopting multi-year assessment periods is not inconsistent with 'annual median' or 'annual 95th percentile' as used in the NOF tables. Under a rolling regime assessment frequency can be annual even though a multi-year assessment period is adopted. Furthermore, rolling assessments pose lower "state switching" frequencies cf. adjacent non-overlapping assessment periods.

One item that will be inconsistent, and probably desirably so, is the future inclusion of sediment aspects into the NOF, where there is some prospect of stating Numeric Attribute States and a percentage change from a reference or upstream value (as in the MfE 1994 Guidelines for water colour and clarity). The statistical implications of such an approach should be considered before finalising the form of these additions. In particular, the feasibility of detecting a percentage change should be examined given the potential for statistical sampling error to frustrate the ability to detect it.

In this context it is helpful to further consider the clarity attribute. Were the reference state to be "declared" (e.g., Wet-Hills clarity bottom line is a median of 1.6 m) we would only be concerned about imprecision from one source of data (the measurement made). But when considering percentage change from an upstream value (sampling on a case by case basis) there are *two* sources of uncertainty—the upstream and downstream datasets, in which case the influence of statistical sampling error will be greater.

The usual approach to these issues uses point-null hypotheses (no difference between the states) and so is subject to the same objections as apply to trend testing. Instead one could adopt the approach advocated for trend testing, inferring the direction of difference (which, after all, is what we want to know). This will offer a more powerful method of detecting change compared to the standard testing of a nil hypothesis (that the distributions' parameters are identical).

5 Human health Value

5.1 Rationale for sample 95%iles for *E. coli*

5.1.1 Assessment period

The *E. coli* table in the National Objectives Framework uses annual medians for secondary contact. So the considerations in section 3.1.2 apply, as *E. coli* in rivers and lakes is expected to follow a lognormal distribution. That is, it may be best to interpret "Annual median" as a rolling annual assessment, using the previous five years of data.

For primary contact ("…undertaking activities likely to involve full immersion…") the 95%ile is used as the assessment metric, and no particular assessment period is stated. So, in this case, it is a simple matter to adopt a rolling five-year assessment period, with an annual assessment frequency. Note however that the MfE/MoH (2003) guidelines do not explicitly call for rolling assessments, though they are contemplated.[38]

5.1.2 Assessment metric

As to the adoption of the 95%ile as the assessment metric, it must be noted that there are two separate paths by which such a criterion may be developed for recreational waters, both depending on the distribution of results of a prior Quantitative Microbial Risk Assessment (QMRA). It all depends on how that distribution is incorporated. First is the adoption of a *distribution descriptor* the second refers to a *precautionary descriptor*.

Distribution descriptor

The distribution descriptor approach was used to develop the coastal water component of the international recreational water quality guidelines (WHO 2003), about which a substantial literature is available: Kay et al. (2004), Wymer et al. (2005), Kay et al. (2006). In this approach a probability density function is identified for the probability of illness, versus concentration of a faecal indicator,[39] based on a series of epidemiological studies. As stated by its major developers (Kay et al. 2004):

> *Using this novel approach, it is possible for the policy community to specify an acceptable excess probability of illness and then to define the parameters of the pdf required (i.e., a geometric mean value, or a 95th percentile value given the knowledge of the standard deviation of log$_{10}$ transformed values) to limit the likely symptom incidence to this level or lower.*

The important point to note is that in this approach a geometric mean (or the equivalent median value[40]) and the 95%ile are <u>different numbers</u>. (Typically, in water pollution studies, the 95%ile is at least twice the median—Ellis 1996—and for a lognormal distribution with a coefficient of variation of 1.5 the ratio is as high as 6:1.[41]) The median and 95%ile are different numbers because one describes central-tendency and the other describes extreme attributes <u>of the same distribution</u>, i.e., the (lognormal) distribution of swimmer's risk of becoming ill, in excess of the risk faced by non-

[38] In MfE/MoH (2003) at page E5, "Step 8: Reassessment" we have: "Reassess on a five-yearly basis, or sooner if significant change occurs. Such changes will be reflected in new information…. Examples of significant change would be: altered catchment characteristics or land use; significantly higher or lower microbiological indicator levels; major infrastructure works affecting water-quality parameters".

[39] The indicator is intestinal enterococci.

[40] For a lognormal population the median and the geometric mean are identical values.

[41] Using parameter equations given by Gilbert (1987) the 95%ile:median ratio is calculated as $\exp\{1.645*\sqrt{\ell n[(1 + \eta^2)]}\}$ where η is the coefficient of variation. Results of this formula agree with examples given by Ellis (1996).

swimmers at the same beach. This is addressed by way of a diagram (Figure 5-1), after first discussing the alternative precautionary descriptor.

Precautionary descriptor

In this approach the median and 95%ile in the freshwater guidelines (MfE/MoH 2003) are the same number, because they refer to different distributions.

To explain this note that a New-Zealand-specific approach to this matter was adopted because the international guidelines do not include explicit numerical guideline values for freshwater. As noted in the equivalent Australian guidelines (NHMRC 2008):

> *It is not possible to directly derive microbial assessment categories for freshwater because of a lack of data.*

It is for that reason that the New Zealand freshwater guidelines (MfE/MoH 2003) were derived from a national QMRA study of water-related campylobacteriosis (this country's major reported notifiable disease), using a dose-response relationship derived from a clinical trial (as described by Till et al. 2004, 2008; McBride 2012). In this approach, 'best guesses' were calculated for the critical value of *E. coli* concentration for given values of excess illness. Essentially the values are medians of a risk profile.[42] So setting these *E. coli* values as median *sample* metrics inherently adopts an 'even-handed' approach to the burden-of-proof. That risk, as always, has to do with statistical sampling error. So, for example, consider what happens if the thresholds were set as annual medians. If the median at a river site was truly stable at 538 per 100 mL (qualifying it as Attribute State B) the chances of obtaining an annual median value greater than 540 per 100 mL (inferring Attribute State C) is about 50%, a high misclassification error risk. Requiring that this threshold be assessed as a 95%ile reduces this risk to about 5%. In effect the risk distribution is then shifted to the left such that its 95%ile now lies at the value of the median before shifting, as depicted in Figure 5-1.

That is, as the lower part of Figure 5-1 shows, a precautionary approach was taken in developing the freshwater component of the 2003 guidelines; not because this mimicked the international approach (for coastal waters) but in order to implement a precautionary stance to the assessments.

Therefore the 95%ile in the WHO (2003) guidelines has a different rationale than their New Zealand equivalent. The former are the more permissive. They are based on the upper part of Figure 5-1.

[42] They were developed using 'percentile matching', as described by McBride (2012, 2014).

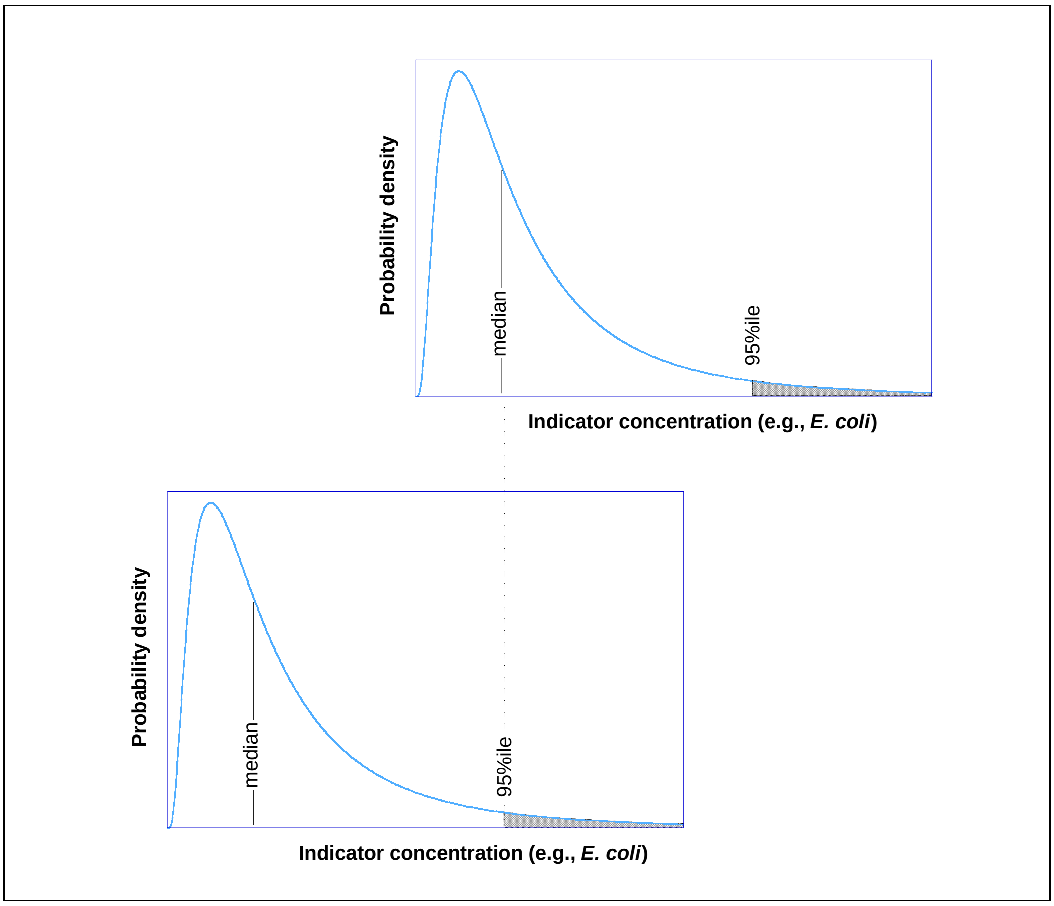


Figure 5-1: Shifted (lognormal) distributions such that the lower distribution has been shifted to the left (a precautionary approach) to make its 95%ile numerically equal to the upper distribution's median. (The cross-hatched right tail in the distribution has an area of 0.05.)

5.2 Accounting for seasonality and flow?

Seasonality applies to both lakes and rivers, but the issue of whether to restrict sampling programmes according to flow arises of course for rivers, rather than lakes.

5.2.1 Secondary contact, rivers

Secondary contact is defined in the National Objectives Framework as "…contact with water during activities with occasional immersion and some ingestion of water (such as wading and boating)." Such activities can take place year round. Even when rivers are in flood, there can be some water contact by people in the proximity, such as may occur from splashing. It seems therefore undesirable that seasonality and flow adjustments be made for this activity.

Furthermore, a major proportion of the load of faecal indicator bacteria can be delivered during flooding events (McKergow and Davies-Colley 2010), so that elevated concentrations of pathogenic material is likely (at least on the rising limb of a flood hydrograph, Stott et al. 2011). Therefore, with one possible caveat, it seems undesirable to exclude seasonality or flood events from monitoring for secondary contact. That exception concerns sampling on the falling limb of a hydrograph. Both McKergow and Davies-Colley (2010, Figure 6) and Stott et al. (2011, Figure 3) present evidence that concentrations of faecal indicator *E. coli* and a pathogen (*Campylobacter*) are elevated on the rising hydrograph limb. However, there is strong hysteresis in that these concentrations on the rising limb at elevated flow are often much higher than the concentrations on the falling limb at the same flow. Given that assessment metric is the median, the presence of these falling limb low concentrations may not adequately reflect the presence of the risk occurring at rising flows when water contact may be more prevalent, given that their duration is typically longer that for the rising limb.

5.2.2 Primary contact, rivers

For primary contact recreation the National Objective Framework's requirements are essentially based on the 2003 water quality guidelines for recreational water contact (MfE/MoH 2003). Those guidelines consist of two strands: site grading and site surveillance. The former is relevant here because it uses sample 95%iles over a five year period, with weekly sampling over the bathing season.

The fundamental requirement of sampling timing is given in section E.2 of MfE/MoH (2003)[43], viz.:

> *Samples should be collected during the bathing season, or when the water body is used for contact recreation. For rivers this may exclude periods of high flow, during which hazardous river conditions would prohibit bathing. The bathing season will vary according to location, but will generally extend from 1 November to 31 March. Sampling should take place between 8 am and 6 pm.*

Therefore sampling can (indeed, should) often be restricted to the bathing season and to non-flood flows. But the manner in which that should be done will vary from one flow regime to another so it is difficult to conceive of a *general* rule that would exclude flood conditions adequately. The same is true for the exposure season; water contact may be happening throughout the year (e.g., boating, primary contact during waka-based cultural water activities (pers. comm. Hannah Rainforth, June 2016).

In addressing this issue, there needs to be a trade-off between loss of precision and the exclusion of elevated flows. That is, if there were no exclusion of elevated flows, then the guidelines' recommended sampling regime would furnish about 100 results, giving good precision in the estimation of a 95%ile. If half of those data were for a flow that exceeded some criterion precision becomes somewhat diminished (as discussed in section 2.2.2).

[43] See also Section H(i) of MfE/MoH (2003) for more detail on sampling times and periods.

6 Conclusions and Recommendations

The most important outcome of these considerations should be a heightened awareness of the pervasive and at-times troublesome role of sampling variability, sometimes denoted as 'statistical sampling error'. In particular, this means that we never know the true Attribute State (as a percentile of time), because we only ever have an estimate of that and such an estimate is influenced by the variability components we happen to capture in our samples. Sometimes it will be higher than the true value, sometimes lower. It will seldom if ever be the true value. Hence the proposed motto: "Always be wary of the influence of 'statistical sampling error'".

For the most part the NOF is silent on the burden-of-proof that underlies the various percentile requirements, seemingly because they have been regarded as percentiles *of samples* rather than percentiles *of time*. If the latter is intended, after all aquatic flora and fauna "see" contamination and its effects *over time*, then there will need to be a direct consideration of the burden-of-proof and misclassification error risks when assessing Numeric Attribute States. The implications for sampling effort are revealed through look-up tables for medians, 80%iles and 95%iles. It is noted that adopting a precautionary approach generally increases considerably the needed number of samples.

Primary contact recreation human health Numeric Attribute State is already based on a precautionary approach to the burden of proof.

The propensity for "State Switching" (e.g., inferring states A-B-A-B in four successive years when in fact the waterbody was always in State B) under annual assessments using only that year's data seems unacceptably high. Instead it is suggested that optimally a five-year assessment period be adopted with rolling annual assessment frequency, with three years as the minimum.

A recently-developed (and implemented) direction-of-trend assessment procedure is recommended for progress assessment for the various attributes, after a few years of data have been analysed.

For the secondary contact in the human health value it is seems largely unnecessary to restrict sampling to seasons and lower flows, with one possible exception. That is, consideration should be given to (if possible) sampling for *E. coli* only on the rising limb of a flood hydrograph. For primary contact there should be sampling stratification based on season and on flow (when conditions may be unsafe for swimming). The identification of what elevated flows and seasons should be so-treated will vary from location to location.

Consistency of approach should be aimed for with future Attributes (or revisions of the current set), but may not always be achievable. Issues that may arise with a percentage change approach for sediment attributes will need careful consideration in the light the findings in this study. In particular, the feasibility of detecting a percentage change should be examined given the potential for statistical sampling error to frustrate the ability to detect it.

6.1 Recommendations

- In order to minimise risks of false state-switching, annual assessments of attribute state attainment using medians should generally be conducted using a rolling window of at least three year's data, preferably five. However adjacent annual assessments may be appropriate for high percentiles (95%ile or maxima).

- In cases where rare events may occur (e.g., lake hypolimnion hypoxia, *E. coli* spikes resulting from a WWTP failure) it may be appropriate to use multi-year rolling assessments for medians and single-year adjacent assessments for 95%iles and maxima.

- Identify the appropriate burden-of-proof for assessing attainment of desired attribute states, between the choices of "precautionary", "permissive" or "even-handed".

- If precautionary or permissive stances are adopted state rules for inferring percentile attainment, use simple already-published "lookup" tables.

- Consider using three-outcome "two one sided" tests for the influence of statistical sampling error on state-switching in which, in the case of A versus B states, there could be three outcomes A, B, or U (undecided).

- For future attributes such as sediment where percentage changes from some reference value is contemplated, the feasibility of detecting a percentage change should be examined given the potential for statistical sampling error to frustrate the ability to detect it.

- For assessing primary contact recreation sites samples should be collected during the bathing season, or when the water body is used for contact recreation. For rivers this may exclude periods of high flow, during which hazardous river conditions would prohibit bathing. But the manner in which that exclusion should be done will vary from one flow regime to another so it is difficult to conceive of a *general* rule that would exclude flood conditions adequately. The same is true for the exposure season; at some sites water contact may be happening throughout the year (e.g., rafting). In addressing this issue, there needs to be a trade-off between loss of precision and the exclusion of elevated flows.

- Changes to the wording of some of the existing Attribute State's assessment metrics should be contemplated. Currently some are described as "Annual Median", "Annual Maximum", "Annual 95th Percentile" or just "95th percentile". For example, "Annual" will generally be interpreted as referring to only one year of data but this analysis has demonstrated that annual assessment using the last three or (better still) five years' data will minimise the role of sampling error, as seen in false "state switching". So the minimum number of samples and minimum duration should be specified. For example, "based on a minimum of a monthly sampling regime and a minimum record length of [e.g., 3 years]".

6.2 A decision template

A three-step consideration is suggested when grappling with these issues.

6.2.1 Choosing a time period and assessment regime

Annual assessments using data collected monthly over one year carries considerable uncertainty and therefore the risk of common false state-switching. Consider using at least three year's data in a rolling window of annual assessments for median thresholds but retain adjacent annual assessments for high percentiles. Also consider the merits of increasing sampling frequency for sensitive water bodies.

6.2.2 Deciding on the burden-of-proof

What burden?	Considerations for policy and decision-making	Attribute considerations
Precautionary		
Prior assumption: Assume NOF threshold has been exceeded until convinced otherwise, e.g., human health protection.	Appropriate when the consequences of wrongly concluding something is 'safe' outweigh the implications of unnecessarily declaring it 'unsafe'. Appropriate if the consequences of the following are too high: • Concluding a popular swimming spot is safe, when in truth there is a high infection risk and people become sick • Concluding the condition of a waterbody is suitable for sustaining a valuable species, when in truth it was not and the species is lost.	While the assessment metric is properly a percentile (of time) its implementation may be most efficient, timely and simple using the look-up tables presented herein.
Permissive		
Prior assumption: Assume a NOF threshold has *not* been exceeded, unless convinced otherwise.	Appropriate when the consequences of wrongly concluding something is 'unsafe' outweigh the implications of mistakenly declaring it 'safe'. Appropriate if the consequences of the following are too high: • Deciding to close a popular swimming spot, when in truth it was safe and could have been used. • Implementing costly mitigation measures on the basis of a valuable freshwater species being harmed, when in truth it was not, and the mitigation was unnecessary.	As above, while the assessment metric is properly a percentile (of time) its implementation may be most efficient, timely and simple using look-up tables.
Even-handed		
Take the data at face-value, making no prior assumption about whether a threshold has been exceeded.	Appropriate when wishing to make attainment assessments 'on the balance of probabilities' (as used in civil law proceedings). But be aware that this stance carries enhanced misclassification risks (because the effects of statistical sampling error have been set aside).	Use direct calculation of sample percentiles, disregarding the pluses and minuses caused by sampling variability.

6.2.3 Choosing a comparison reference?

Make an explicit choice about whether any new Numeric Attribute States (such as sediments) should be described relative to an environmental condition (e.g., an instream value) versus a reference or control state, bearing in mind that the latter option will possess less inherent statistical sampling error.

6.3 Examples using the template

Three examples are described below: total phosphorus in lakes; *E. coli* in rivers; visual clarity in rivers.

6.3.1 Total phosphorus attribute table for lakes

1. *Time period and assessment regime*
   This report finds that an annual assessment using only one year of data (e.g., sampled monthly as may be typical for lakes) does not confer satisfactory precision of attainment of attribute state, assessed using sample medians. At least three years of monthly data should be included in a rolling assessment regime, to minimise the occurrence of false state-switching. Using adjacent assessments seems inappropriate in this case even if large spikes in TP are thought to be rare: they would occur as the result of long-term processes and so a large TP spike is suggestive of future spikes occurring more frequently.

2. *Burden of proof*
   Decisions about the acceptable or tolerable degree of precaution or permissiveness depend on the communities' or decision makers' tolerance to the risks and tradeoffs. For example, unlike *chlorophyll a*, spikes in lake TP are not obvious to the eye. If we were to incorrectly conclude that TP levels are acceptable, when in reality they were not, we would be unaware of impacts on lake ecology and would not undertake a management response. The implication is that the problem would become advanced before it would be detected, and may be expensive to address (if possible). If we assume this is unacceptable, a precautionary approach should be taken for the implementation of the NOF's TP Table (the NOF's TP table's formulation was based on an even-handed stance, cf. precautionary, so there would be no precautionary 'double-up'). Instead of calculating the median of the dataset and comparing that with the "Annual Median" thresholds in the NOF TP table, the appropriate look-up table should be used (Table 2-3). So for 36 samples, in order to be assessed as Attribute State A only 13 of them can exceed a TP concentration of 10 mg/m^3, i.e., 36%. Were an even handed stance to be taken, 18 samples could exceed that threshold. Similar considerations arise for the other thresholds in the NOF Attribute table.

3. *Choosing a comparison reference?*
   Not relevant in this case.

6.3.2 *E. coli* attribute table for rivers

1. *Time period and assessment regime*
   Annual assessment using only one year of data (sampling monthly or weekly during a defined bathing season) does not confer satisfactory precision of attainment of Attribute state. This situation is made even worse when some intended sampling dates are abandoned (because of high flows). Annual rolling assessment over the most recent three to five years of data is preferable, for bother medians (secondary contact) and 95%iles (primary contact).

2. *Burden of proof*
   As discussed in the text, a precautionary approach to the burden-of-proof was taken in the formulation of the MfE/MoH 2003 Guidelines, and therefore the NOF *E. coli* Attribute table. This was achieved by setting the table thresholds as 95%iles (see section 5.1.2). It is therefore undesirable to double-up this precaution by also adopting a precautionary approach in the implementation of the NOF Table. It follows that the sample 95%ile (from the three or five years of data, using the Hazen calculation method) could be calculated at face-value and compared with thresholds. Note that it may be more desirable (and understandable to the public) to simply state that up to 5% of the samples can exceed the 95%ile threshold value (Hunter 2002).

3. *Choosing a comparison reference?*
   Not relevant in this case.

6.3.3 Future visual clarity attribute table for rivers?

1. *Time period and assessment regime*
   It is likely that there will be both median and high percentile thresholds, as there is for nitrate in rivers, in which case rolling annual assessments, based on three to five years data, could be used for the median. Adjacent annual assessments should be contemplated for the high percentile. There may need to be some consideration of whether assessment of attribute states should be restricted

2. *Burden of proof*
   In contrast to the TP example above, changes in visual clarity are (by definition!) obvious to the eye. Therefore it seems more appropriate to use an even-handed approach to the both the setting of the thresholds and the subsequent assessments of attainment. Note that this bears on the question of whether the high percentile should be set as a 95%ile or as a maximum (i.e., a 100%ile). If sampling is monthly an even-handed approach could be argued to permit one exceedance per year (if 12 samples are obtained).[44] Certainly one exceedance could be allowed under fortnightly sampling. But no exceedances of a 100%ile can be allowed, regardless of sampling frequency.[45]

3. *Choosing a comparison reference*
   If the comparison is to a 'gold standard' the influence of statistical sampling error will be confined to the sampling effort at the site for which an assessment is to be made. But if the comparison is to be for a change from upstream conditions (as in 'Guideline 1' in MfE 1994) that variance will be inflated because natural variability at both upstream and downstream sites will need to be considered. This increases uncertainty in assessments. The degree to which this is important could be considered as a separate study.

[44] That is because 5% of 12 is 0.6 which is closer to 1 than it is to zero. But if restrictions to lower flows only are imposed there would be less than 12 samples and so the case for permitting one exceedance per year is weakened.

[45] It should be noted that the nitrate (toxicity) table has thresholds for annual 95%ile whereas the ammonia (toxicity) table has thresholds for annual maxima. Under a precautionary approach both would forbid any exceedances of their thresholds but, as noted above, under an even-handed approach arguably one exceedance per year of a 95%ile threshold (under monthly sampling) could be entertained (and under a permissive approach two exceedances could be allowed—see Table 2-4.)

7 References

Barnett, V., O'Hagan, A. (1997) *Setting Environmental Standards: The Statistical Approach to Handling Uncertainty and Variation*. Chapman and Hall, London.

Bross, I.D. (1985) Why proof of safety is much more difficult than proof of hazard. *Biometrics*, 41: 785–793.

Burns, N.M., Rutherford, J.C. (1998) *Results of monitoring New Zealand lakes 1992–1996*. NIWA Client Report MFE80216, to Ministry for the Environment.

Cohen, J. (1994) The earth is round (p<05). *American Psychologist*, 49(12): 997–1003.

Conover, W.J. (1980) *Practical Nonparametric Statistics*. 2nd ed. Wiley, New York.

Davies-Colley, R.J., Hughes, A.O., Verburg, P., Storey, R. (2012) Freshwater monitoring protocols and quality assurance (QA): National Environmental Monitoring and Reporting (NEMaR) variables. *NIWA Client Report Ham2012-092* to Ministry for the Environment, Project MFE12201. 104 p.

Ellis, J.C. (1986) *Determination of pollutants in effluents. Part B. Alternative forms of effluent consents: some statistical considerations*. Report TR 235, Water Research Centre, Medmenham, England, April.

Ellis, J.C. (1989) *Handbook on the design and interpretation of monitoring programmes*. Report NS 29, Water Research Centre, Medmenham, England (first published April 1990).

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

Hunter, P.R. (2002) Does calculation of the 95th percentile of microbiological results offer any advantage over percentage exceedence in determining compliance with bathing water quality standards? *Letters in Applied Microbiology*, 34: 283–286.

Ihaka, R., Gentleman, R. (1996) R: A language for data analysis and graphics. *Journal of Computational and Computational Statistics*, 5(3): 299–314.

Jones, L.V., Tukey, J.W. (2000) A sensible formulation of the significance test. *Psychological Methods*, 5(4): 411–414.

Kay, D., Bartram, J., Pruss, A., Ashbolt, N., Wyer, M.D., Fleisher, J.M., Fewtrell, L., Rogers, A., Rees, G. (2004) Derivation of numerical values for the World Health Organization guidelines for recreational waters. *Water Research*, 38: 1296–1304.

Kay, D., Ashbolt, N., Wyer, M.D., Fleisher, J.M., Fewtrell, L., Rogers, A., Rees, G. (2006) Reply to comments on "Derivation of numerical values for the World Health Organization guidelines for recreational waters". *Water Research*, 40: 1921–1925.

Larned, S., Snelder, T., Unwin, M., McBride, G., Verberg, P., McMillan, H. (2015) *Analysis of Water Quality in New Zealand Lakes and Rivers*. Prepared for Ministry for the Environment, *NIWA Client Report CHC2015-033*, Project MFE15503, 74 p. plus Appendices A and B.[46]

[46] https://data.mfe.govt.nz/documents/category/environmental-reporting/freshwater/river-water-quality/

Larned, S., Snelder, T., Unwin, M., McBride, G.B. (2016) Water quality state and trends in New Zealand rivers, 2004–2013. *New Zealand Journal of Marine and Freshwater Research.*[47]

McBride, G.B. (2003) Confidence of compliance: parametric versus nonparametric approaches. *Water Research*, 37(15): 3666–3671

McBride, G.B. (2005) *Using Statistical Methods for Water Quality Management: Issues, Problems and Solutions*. John Wiley & Sons, New York.

McBride, G.B. (2012) *Issues in setting secondary contact recreation guidelines for New Zealand freshwaters*. Report to the Ministry for the Environment, Wellington, 9 September: 12 p.

McBride, G.B. (2014) *National Objectives Framework for Freshwater: Statistical considerations for assessing progress towards objectives with emphasis on secondary contact recreation values NIWA Client Report No: HAM2014-007*, Prepared for Ministry for the Environment, Project MFE14202, 32 p. February.

McBride, G.B., Ellis, J.C. (2001) Confidence of Compliance: a Bayesian approach for percentile standards. *Water Research*, 35(5): 1117–1124.

McBride, G.B., Till, D., Ryan, T., Ball, A., Lewis, G., Palmer, S., Weinstein, P. (2002) *Freshwater Microbiology Research Programme. Pathogen Occurrence and Human Health Risk Assessment Analysis*. Ministry for the Environment Technical Publication: 93 p. http://www.mfe.govt.nz/publications/water/freshwater-microbiology-nov02/

McBride, G.B., Cole, R., Westbrooke, I., Jowett, I.G. (2014) Assessing environmentally significant effects: A better strength-of-evidence than a single *P* value? *Environmental Monitoring and Assessment*, 186(5): 2729–2740 doi: 10.1007/s10661-013-3574-8.

McBride, G., Snelder, T., Unwin, M., Booker, D., Verburg, P., Larned, S. (2015) A new approach to water quality trend assessment. Appendix A in Larned, S.; Snelder, T.; Unwin, M.; McBride, G.; Verburg, P.; McMillan, H. Analysis of Water Quality in New Zealand Lakes and Rivers. Prepared for Ministry for the Environment, *NIWA Client Report CHC2015-033*, Project MFE15503, 74 p. plus Appendices.[48]

McGrayne, S.B. (2011) *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*. Yale University Press, MA.

McKergow, L.A., Davies-Colley, R.J. (2010) Stormflow dynamics and loads of Escherichia coli in a large mixed land use catchment. *Hydrological Processes,* 24: 276–289.

MfE (1994) Resource Management Water Quality Guidelines No. 2: Guidelines of the Management of Water Colour and Clarity, Ministry for the Environment, 77 p.

Millard, S.P. (1998) *Environmental Stats for S-Plus: User's Manual for Windows® and Unix®.* Springer, New York.

Millard, S.P., Neerchal, N.K. (2001) *Environmental Statistics with S-Plus.* CRC Press, Boca Raton, FL.

[47] http://dx.doi.org/10.1080/00288330.2016.1150309
[48] https://data.mfe.govt.nz/documents/category/environmental-reporting/freshwater/river-water-quality/

MfE/MoH (2003) *Microbiological Water Quality Guidelines for Marine and Freshwater Recreational Areas*. Ministry for the Environment and Ministry of Health, Wellington, New Zealand. (http://www.mfe.govt.nz/publications/water/microbiological-quality-jun03/)

Mood, A.M., Graybill, F.A. (1963) *Introduction to the Theory of Statistics*. 2nd ed., McGraw-Hill, New York.

NZGovernment (2014) *National Policy Statement for Freshwater Management 2014*. Issued by notice in gazette on 4 July 2014. http://www.mfe.govt.nz/publications/fresh-water/national-policy-statement-freshwater-management-2014

NHMRC (2008) *Guidelines for Managing Risks in Recreational Water*. Australian Government / National Health and Medical Research Council, Canberra. (https://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/eh38.pdf)

Norton, N., Harris, S., Rouse, H. (2015) *Guidance on communicating and managing uncertainty when implementing the National Policy Statement for Freshwater Management (2014).* Land Water People Ltd., *NIWA draft report to Ministry for the Environment*, version 3, 30 October 2015.

Reckhow, K.H., Chapra, S.C. (1983) *Engineering Approaches for Lake Management. Volume 1: Data Analysis and Empirical Modelling*. Butterworth, Boston.

Stott, R., Davies-Colley, R., Nagels, J., Donnison, A., Ross, C., Muirhead, R. (2011) Differential behaviour of *Escherichia coli* and *Campylobacter* spp. in a stream draining dairy pasture. *Journal of Water and Health*, 9(1): 59–69.

Till, D., McBride, G.B. (2004) Potential public health risk of *Campylobacter* and other zoonotic waterborne infections in New Zealand. Chapter 12 in *Waterborne Zoonoses: Identification, Causes and Control.* Cotruvo, J.A., Dufour, A., Rees, G., Bartram, J., Carr, R., Cliver, D.O., Craun, G.F., Fayer, R., Gannon, V.P.J., eds. World Health Organization (WHO). IWA Publishing: London, UK.

Till, D., McBride, G., Ball, A., Taylor, K., Pyle, E. (2008) Large-scale freshwater microbiological study: Rationale, results and risks. *Journal of Water Health*, 6: 443–460.

Ward, R.C., Loftis, J.C., McBride, G.B. (1990) *Design of Water Quality Monitoring Systems*. Van Nostrand Reinhold, New York. 231 p.

Wymer, L.J., Dufour, A.P., Caldron, R.L., Wade, T.J., Beach, M. (2005) Comment on “Derivation of numerical values for the World Health Organization guidelines for recreational waters”. *Water Research*, 39: 2774–2777.

WHO (2003) *Guidelines for Safe Recreational Water Environments. Volume 1: Coastal and Fresh Waters*. World Health Organization, Geneva. http://www.who.int/water_sanitation_health/bathing/srwe1/en/

8 Acknowledgements

Carl Howarth (especially), Vera Power and Lauren Long (Ministry for the Environment), Ton Snelder (Land Water People) and Jeff Soller (Soller Environmental, USA, CA) clarified the issues to be addressed. Neale Hudson (NIWA) reviewed the whole report; James Sukias (NIWA) also reviewed chapters 3 and 5. Chris Hickey (NIWA) made useful suggestions for section 3.2.

9 Glossary of abbreviations and terms

Accuracy	High precision, low bias.
Assessment frequency	The time between adjacent assessments, typically one year.
Assessment metric	A sample statistic, such as a median or a 95%ile.
Assessment period	The number of years of data to be included in each assessment.
Bias	A tendency to be ‘off the mark’.
Confidence interval	Ranges within which a parameter may lie most of the time, under repetitive sampling.
Even-handed	Taking the sample estimate as the true population value, ignoring statistical sampling error. Same as “face-value”.
Percentile	Same as “quartile” or “centile”. A value below which a given percentage of data fall. Can refer to either samples of populations.
Precision	Lack of scatter of estimates about a true value.
Probability density function (pdf)	The shape of a histogram were an infinite number of samples to be taken and measured accurately. It is not a probability; probabilities are areas under pdf (the total area under a pdf is 1).
Progress assessment	Tracking movement of an attribute’s sample statistic(s) toward (or away from) a desired attribute state.
Proof of hazard	A testing procedure that assumes the hazard does not exist which is only rejected if new data strongly indicate to the contrary. Also called the *permissive approach*, “slipping through the net”, “letting the guilty go free”, or “benefit of doubt”.
Proof of safety	A testing procedure that assumes the hazard does exist which is only rejected if new data strongly indicate to the contrary. Also called the *precautionary approach*, or “fail-safe”.
QMRA	Quantitative Microbial Risk Assessment, in calculations are made from data or assumptions concerning human exposure to pathogens (or their indicators) in water, calculating risk profiles from dose-response relationships.
Statistical sampling error	The difference between a sample statistic used to estimate a population parameter and the actual, but unknown, value of that parameter. Here “error” does not imply that there has been a mistake; it is a technical term in statistical parlance relating to accuracy.
Tolerance interval (one-sided)	A percentile inflated or deflated a little to take account of statistical sampling error.

Appendix A Interpreting confidence intervals

Strictly, a confidence interval with numeric limits should be interpreted in a frequency sense, the so-called 'classical perspective'.[49] In particular, if an analyst were to conduct a sampling effort many times and on each occasion computed a 95% confidence interval, then on average 95% of those intervals would contain the estimated parameter (Mood and Graybill 1963). In other words, the frequency approach to probability only allows us to calculate probabilities of obtaining a given range of data, under repetitive sampling.

These considerations are depicted in the following figure (guided by Mood & Graybill 1963, p. 253). In this figure ten 90% confidence intervals for an estimate of the mean value of a water quality variable are depicted, one of which (the blue line) does not contain the true value.[50] This figure depicts the 'correct' frequentist interpretation of a confidence interval.

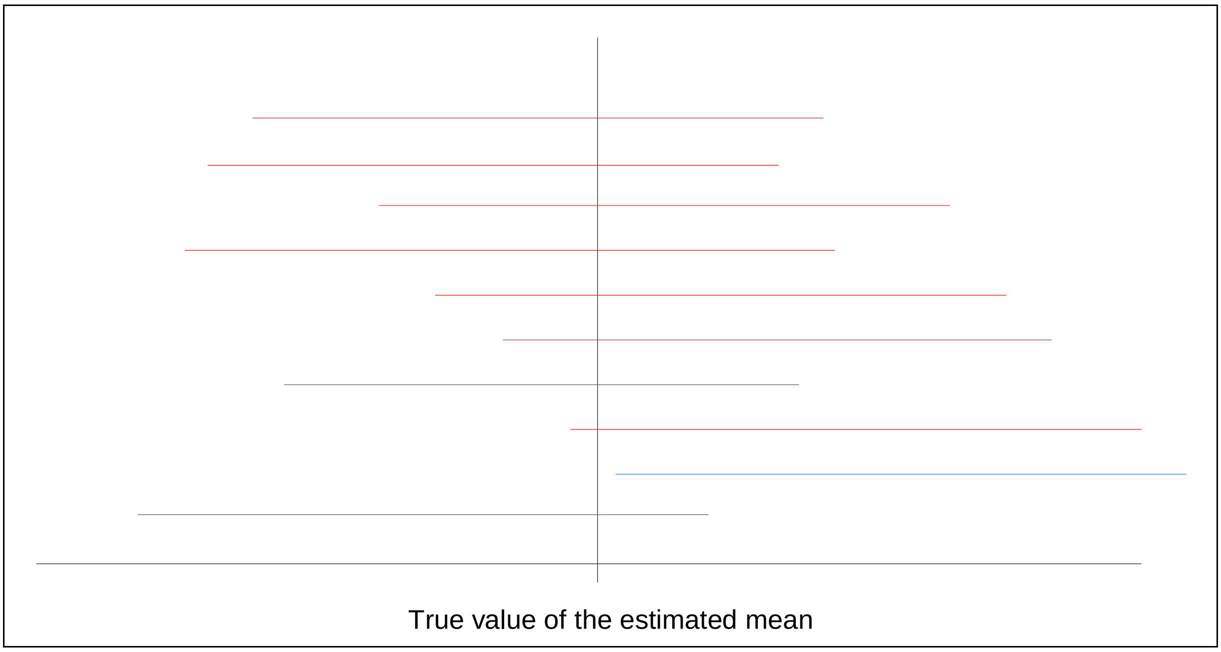


Figure A-1: Ten 90% confidence intervals for estimating the mean where the blue line does not include the true value but the red lines do.

However most workers interpret a single confidence interval as confidence that the interval actually obtained contains the true value of the parameter. That's not surprising, given that no-one has the resources to gather multiple datasets (and the population will be changing over time). But one should be aware that this interpretation invokes the **Bayesian** view of probability: expressing a 'prior' degree of belief about the parameter (Reckhow & Chapra 1983, p. 76). These authors have noted that:

[49] 'Classical' is something of a misnomer given that it only really emerged in the early 20th century whereas the earlier (Bayesian) approach dates back to the 18th century.

[50] Of course sometimes more than one of these intervals would not contain the true mean, sometimes none would; that's the vagaries of statistical sampling error!

> *It is interesting that most researchers are taught statistics from a classical perspective, yet confidence intervals are often interpreted in a Bayesian sense. When the Bayesian interpretation is adopted, the analyst should realize that this implies a subjective interpretation for probability, and this should be specified in the analysis … the prior probability distribution must be stipulated if the Bayesian interpretation for confidence intervals is adopted….*

Few heed this advice, but more may be expected to do so in years to come, especially with the advance of Bayesian software (including freeware, such as the R package).[51]

Note that in a Bayesian interpretation, new data are used to modify prior distributions into posterior distributions, using **Bayes rule**. The rule itself is not controversial but its application to confidence intervals and hypothesis testing is (McGrayne 2011). That's because different analysts will often hold different degrees of belief before considering new data, and so their results (as posterior intervals and distributions) will diverge, particularly for small datasets. Referring back to confidence intervals (section 2.1.6) we observe that the prior belief unwittingly invoked when making their common Bayesian interpretation is generally "non-informative", in that it posits that all possible values of the parameter are equally likely: it is "flat" or "vague". In many cases this could be considered too extreme and the prior distribution could be given some shape whereby some parts of the possible data range are considered to be more likely than others. In that case the resulting Bayesian confidence interval—called a **credible interval**"—will be narrowed. This possibility is not pursued here, but seems worthy of more detailed consideration in the future.

[51] 'R' is a Kiwi initiative, now implemented world-wide. It was developed at the University of Auckland (Ihaka & Gentleman 1996).

Appendix B Tolerance intervals

As noted in section 2.1.8, a tolerance interval limit is effectively a percentile inflated or deflated a little to take account of statistical sampling error.

To be more precise, tolerance intervals are ranges covering a stated proportion of the population most of the time, under repetitive sampling. In particular, they are intervals in which, with a stated confidence level, a specified proportion of a sampled population falls (the proportion is denoted as β). We commonly use "β-content" intervals, constructed so that they contain *at least* 100β% of the population, with a given confidence.[52] The interval's "coverage" is 100β%. Like confidence intervals they can be one-sided or two-sided. See below for an example.

If used in NOF assessments these intervals should be taken as one-sided, because our interest is whether or not a breakpoint has been exceeded (cf. two-sided intervals). One-sided β-content tolerance intervals are calculated from the same formula as is used for one-sided confidence intervals[53]—but that is not the case for two-sided intervals (Conover 1980, Millard & Neerchal 2001, McBride 2005).

In contrast to confidence intervals, two-sided tolerance intervals do not shrink to zero width at infinite sample size. One-sided tolerance limits shrink only to the percentile value.

Example

For a year of monthly *E. coli* sampling, say we had these twelve results: 450, 220, 124, 222, 421, 1020, 311, 222, 222, 355, 622, 490 *E. coli* per 100 mL. Using formulae presented by McBride (2014, Appendix D) we calculate an upper one-sided 95% tolerance limit as 1595 per 100 mL, whereas the Hazen formula for direct calculation of the 95%ile is 980 per 100 mL. But were these data to be repeated exactly for each of the next four years (so the sample size increases to 60), the Hazen percentile estimate rises slightly to 1020 per 100 mL but the tolerance limit reduces to almost exactly the Hazen result (it is 1019 per 100 mL). This example demonstrates that the tolerance limit shrinks to the appropriate percentile as the sample size increases, because there is less uncertainty at larger sample size.

[52] The alternative "β-expectation" intervals are constructed so that they contain *on average* 100β% of the population, with a given confidence. These are not appropriate in this context.

[53] The probability statement for an upper one-sided *confidence* limit (*X*) is: Prob(the β percentile ≤ *X*) = 1 – α, where "1 – α" is the confidence level. Although not straightforward, this equation can be solved for *X*. The probability statement for an upper one-sided β-content *tolerance* limit is: Prob(at least a percentage β of a population ≤ *X*) = 1 – α. It too can be solved for the same value, i.e., *X*. As noted by Conover (1980, p. 120), "These two statements are merely different ways of stating the same idea."