INTERNATIONAL STANDARD

ISO 5667-20

First edition
2008-03-15

Water quality — Sampling —

Part 20: Guidance on the use of sampling data for decision making — Compliance with thresholds and classification systems

Qualité de l'eau — Échantillonnage —

Partie 20: Lignes directrices relatives à l'utilisation des données d'échantillonnage pour la prise de décision — Conformité avec les limites et systèmes de classification

Reference number
ISO 5667-20:2008(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO 5667-20 was prepared by Technical Committee ISO/TC 147, *Water quality*, Subcommittee SC 6, *Sampling (general methods)*.

ISO 5667 consists of the following parts, under the general title *Water quality — Sampling*:

— *Part 1: Guidance on the design of sampling programmes and sampling techniques*

— *Part 3: Guidance on the preservation and handling of water samples*

— *Part 4: Guidance on sampling from lakes, natural and man-made*

— *Part 5: Guidance on sampling of drinking water from treatment works and piped distribution systems*

— *Part 6: Guidance on sampling of rivers and streams*

— *Part 7: Guidance on sampling of water and steam in boiler plants*

— *Part 8: Guidance on the sampling of wet deposition*

— *Part 9: Guidance on sampling from marine waters*

— *Part 10: Guidance on sampling of waste waters*

— *Part 11: Guidance on sampling of groundwaters*

— *Part 12: Guidance on sampling of bottom sediments*

— *Part 13: Guidance on sampling of sludges from sewage and water treatment works*

— *Part 14: Guidance on quality assurance of environmental water sampling and handling*

— *Part 15: Guidance on preservation and handling of sludge and sediment samples*

— *Part 16: Guidance on biotesting of samples*

— *Part 17: Guidance on sampling of bulk suspended solids*

— *Part 18: Guidance on sampling of groundwater at contaminated sites*

— *Part 19: Guidance on sampling of marine sediments*

— *Part 20: Guidance on the use of sampling data for decision making — Compliance with thresholds and classification systems*

The following parts are under preparation:

— *Part 21: Guidance on sampling of drinking water distributed by non-continuous, non-conventional means*

— *Part 22: Guidance on design and installation of groundwater sample points*

— *Part 23: Determination of significant pollutants in surface waters using passive sampling*

Introduction

This part of ISO 5667 concerns the use of information on water quality obtained by taking samples in taking decisions — in measuring success, failure or change, in the context of the inevitable uncertainties associated with sampling. This part of ISO 5667 provides guidance on controlling the risk of such uncertainties leading to non-optimal decisions.

Non-optimal decisions can also stem from the way in which thresholds for discharges and targets for environmental waters are formulated or set out in regulations and permits. This part of ISO 5667 also examines the problems caused when compliance with these thresholds is assessed using data obtained by sampling.

This part of ISO 5667 aims to ensure that future laws, regulations, and guidance assert the requirement to assess and report statistical significance.

NOTE 1 Decisions might result in the commendation or criticism of people, sites, companies, sectors or nations. Decisions can give rise to legal action and/or expensive and time-consuming remedial actions to improve water quality.

Figure 1 shows the links between the following topics:

a) the setting up of thresholds for taking decisions on the need to improve water quality, possibly including criteria to minimize water quality deterioration;

b) the establishment of sampling programmes to satisfy the requirements of these thresholds and the need to assess performance against them;

c) making use of the outcome of sampling programmes to take decisions.

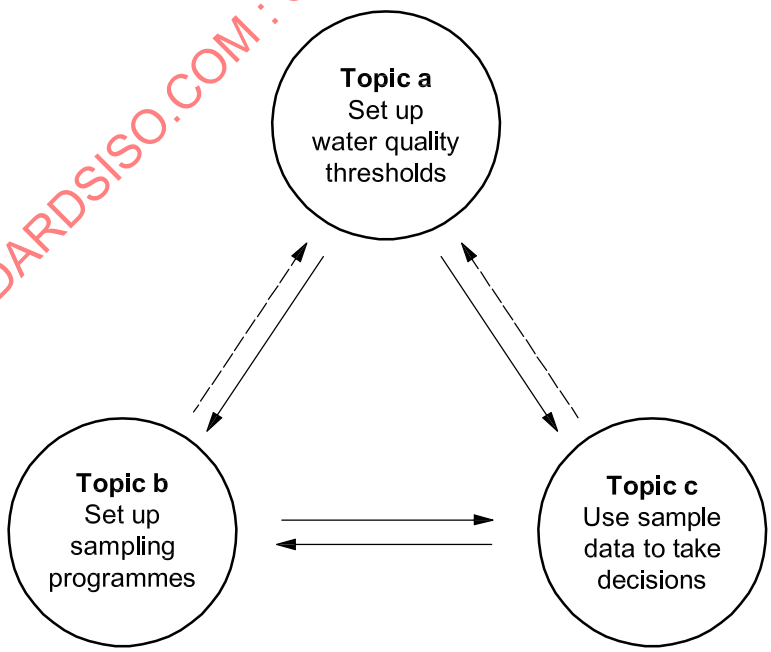


Figure 1 — Links between topics associated with sampling and taking decisions

This part of ISO 5667 deals with topic c). Topics a) and b) are huge and wide ranging in their own right, and their detailed treatment lies outside the scope of this part of ISO 5667. Nevertheless, this part of ISO 5667 does make recommendations for the expression of targets and thresholds for water quality [topic a)], which are important when using sample data to take decisions. This part of ISO 5667 also gives advice on what is required for sampling programmes [topic b)] in order that they be compatible with the way thresholds are defined, and so as to place no unnecessary difficulties and errors in the process of taking decisions.

Other areas which lie outside the scope of this part of ISO 5667 are: the detailed mechanics of taking and handling samples; assurance that samples are representative over time of the body of water being sampled; and performance of chemical analyses on samples. These are all covered in other documents. Nonetheless, if poorly obtained results from these areas can add substantially to overall sampling uncertainties and impose extra difficulties in taking decisions. This part of ISO 5667 describes some of these extra difficulties.

This part of ISO 5667 does not cover the full range of statistical techniques that may be applied and the circumstances in which they should be used. The main purpose is to establish the principle that uncertainty from sampling and analysis (and errors generally) should always be assessed and taken into account as part of the process of taking decisions. If this is not done, incorrect decisions can result, for example, on where action is needed, and the scale of that action.

NOTE 2 Some statistical techniques are used as illustrative examples. These are techniques that have seen routine use in some regulatory regimes that take proper account of statistical uncertainties. They are suitable for use in situations that resemble the worked examples discussed.

It is not the purpose of this part of ISO 5667 to direct the development of regulatory conditions. This part of ISO 5667 provides principles and tools to support management, including regulation. It is recognised that regulatory thresholds are developed using a range of strategies that incorporate technical, social and legal considerations. It is also recognised that tools other than statistical data analysis are likely to be used in interpreting and applying thresholds.

Water quality — Sampling —

Part 20: Guidance on the use of sampling data for decision making — Compliance with thresholds and classification systems

1 Scope

This part of ISO 5667 establishes principles, basic requirements, and illustrative methods for dealing with the use of sample data for decision making based on the assessment of the confidence that water quality:

a) meets targets and complies with thresholds;

b) has changed; and/or

c) lies in a particular grade in a classification system.

This part of ISO 5667 also specifies methods for preliminary examination of the sensitivity of decisions to error and uncertainty, although it does not cover the full range of statistical techniques.

This part of ISO 5667 provides general advice on decision making related to constraint formulation for expression of thresholds and targets and the form and scale of sampling programmes.

NOTE 1 In the water industry, “standard” is commonly used to indicate the value or limit of a parameter of interest. However, in this part of ISO 5667, the term “threshold” is used to avoid confusion with published national, regional, and International Standards.

NOTE 2 This document is framed in terms of sampling and measurement of chemical concentrations, in particular those subject to strong day-to-day temporal variations. The principles apply, however, to any item estimated by sampling which is subject to random error, including microbiological and biological data, and data subject to strong spatial variations.

2 Summary of key points

Water quality is often assessed by the results of chemical analysis of a number of samples taken over a period of time.

Uncertainty is introduced by the action of random chance in taking samples. It can be present in any set of measurements of water quality taken over a period of time. The values for chemical analysis of these samples depend on the quality of the particular small volumes of water that are extracted or measured. If water quality varies in space or time, a second set of samples taken over the same period will have different values because these samples are made up of different small volumes of water taken at different times. Each set of samples allows an estimate of the true water quality. These estimates will differ: they will have a different mean and span a different range. They have the potential, if taken at face value, to suggest different conclusions about compliance with thresholds and targets.

Sampling uncertainty (or sampling error) is the term often given to this effect. Sampling uncertainty includes uncertainties and errors associated with chemical analysis, and occurs even in the case of trivial errors in chemical analysis and if there are no mistakes in the methods by which samples are taken and handled.

Sampling uncertainty is reduced if more samples are taken, but the scale of the uncertainty is often unappreciated.

In this part of ISO 5667, "overall uncertainty" includes these chance sampling effects and all the other sources of variation in a set of samples. This variability reflects the underlying signals generated by natural or perhaps unnatural processes; it includes the effects of errors in chemical analysis and the handling of samples. It might contain systematic variations from trends and diurnal, weekly, and seasonal cycles. In this context, the more appropriate term is "overall uncertainty", "overall error" or "total assay error" (ISO/IEC Guide 99:1993[5]).

Overall uncertainty should be quantified, at least approximately, and taken into account in all cases where water quality varies and sampling is used to estimate information used in decision making. This includes assessing compliance with thresholds (see Clause 5), deciding whether water quality has changed (see Clause 7), and putting waters into grades in classification systems (see Clause 8). This part of ISO 5667 recommends that:

a) thresholds for which compliance is assessed by sampling should be defined or used so that the overall uncertainty can be estimated and dealt with appropriately (see 5.2);

b) thresholds defined as absolute limits should be treated as percentiles when assessing compliance using sampling (see 5.3);

c) thresholds defined as limits to be met by a percentage of samples should be defined or used as the corresponding percentiles (see 5.4);

d) the degree of confidence should be estimated when assessing compliance with thresholds (see Clause 4); and,

e) the degree of confidence in changes or differences should be estimated when aiming to demonstrate change or no change (see 8.2).

3 Types of error and variation

3.1 General

In many procedures by which sample data are used to take decisions, there is a set of results taken over a period of time (e.g. a year). This information might be used to make such judgements as whether:

a) water quality in a river failed to meet required thresholds;

b) a treatment works performed better this year than last;

c) water quality in a lake needs improvement;

d) one company has better effluent discharge compliance than another; or

e) most of the risk of environmental impact is from a particular type of effluent discharge.

There are unlikely to be many significant changes in water quality from second to second throughout a year, but variations from day to day are common. These can be due to diurnal cycles, the play of random errors and bias from the laboratory, the weather, step changes, day-to-day and hour-by-hour variations (perhaps in the natural processes in water or caused by discharges and abstractions and changes in these), seasonal and economic cycles, and several underlying and overlapping long-term trends and cycles.

NOTE 1 Sometimes several or most of the data are reported by a laboratory as being less than a specified limit of detection. Such data are called censored data. Depending on the types of decisions that depend on the data, special statistical techniques are available for estimating the values of summary statistics and their uncertainties.

In addition, the total set of samples shall be representative of the average quality of the masses of water from which they were taken, e.g. over a period of time under review. In estimating an annual mean, it is not acceptable for all samples to be taken in April, for example. These requirements should be set up in the design of the sampling programme.

NOTE 2 Guidance on all these aspects is given in more detail in ISO 5667-1[1].

3.2 Analytical error

Analytical errors are those introduced by the process of chemical analysis and reflect that these measurements are not error free. It might be that the result for a single sample can be specified to within a specific range, e.g. ±15 %.

NOTE 1 The actual value of the analytical error depends on the capabilities of the equipment and the laboratory that has been used to perform the analysis. The discussion in this part of ISO 5667 focuses on random error, but there is always a risk of non-random error, e.g. when there is a change of instrument or methods, when the sample matrix varies greatly from the calibration materials, and for results just above the detection limit (ISO/IEC Guide 99:1993[5]).

NOTE 2 The results of chemical analysis are nowadays reported with uncertainty values in accordance with ISO/IEC 17025[6].

When a mean is calculated from n samples, the effect on the uncertainty in the estimate of the mean of random errors in chemical analysis tends to average down according to $\sqrt{n}$. For example, if the analytical error associated with a single sample were ±15 %, then the error in the estimate of the mean of a set of chemical analyses would tend to reduce to something like ±4 % for 12 samples or to ±2,5 % for 36 samples.

In using samples to take decisions, this kind of error from chemical analysis augments but is often smaller than other contributions to the overall uncertainty, especially that associated with chance in the taking of a limited number of samples. Chemical analysis error comes through as an addition to that associated with chance in the taking of a limited number of samples, but it might be a small addition. {Nevertheless, some studies need to separate sampling variance from local environmental heterogeneity (see Reference [7]).}

NOTE 3 It is not commonly understood that data fully within the statistical control of a laboratory might be unsuitable for particular interpretations because of errors associated with taking a small number of samples.

NOTE 4 This observation on the relative importance of analytical error applies generally to the types of issues considered in this part of ISO 5667, but it follows from estimating the analytical error in such cases, and comparing it with other errors. The analytical error should always be estimated. Similar points can be made about making sure samples are representative, and about checking changes to methods of sampling.

When the sample results are used to estimate the value of other summary statistics such as percentiles (e.g. the 95-percentile, which is the value exceeded for 5 % of the time), the picture is similar to that for the mean, i.e. the errors are inversely proportional to $\sqrt{n}$, but are larger than for the mean.

3.3 Overall uncertainty

Uncertainty occurs because of variations in the quality of the water being sampled, and the ability of the sampling process to accurately reflect these variations. In a set of samples taken over a period of time, the results are affected by the operation of the laws of chance in the way the particular samples came to be collected. This produces uncertainty even if:

— analytical errors are close to zero[1)];

— the sampling programme guarantees samples that are truly representative in time and space;

— there are no mistakes in handling the samples and recording the results of analysis.

1) Nearly always this is a hypothetical possibility. Many trace elements are measured near their detection limits and have analytical uncertainty of about ±100 %. Many organic chemicals can have recoveries of ±50 %.

In using sampling, the main source of uncertainty is usually associated with the number of samples taken. In the types of decision on activities listed in items a) to f) below, this source of uncertainty is usually a bigger issue than, for example, that associated with errors of chemical analysis. Overall uncertainty should be assessed and used to quantify uncertainty in cases where water quality varies and decisions are taken as a consequence of the following types of activities:

a) using sampling to measure and report on water quality;

b) using samples to estimate summary statistics, e.g. the monthly mean, the annual percentile or the annual maximum;

c) making statements about whether this year's summary statistics are higher or lower than last year's (see ISO/IEC 17025[6] for a wider view of the issue of looking for change);

d) establishing whether water quality exceeds a threshold;

e) using summary statistics to place water quality in a particular class within a classification system; or

f) assessing whether a change in class has occurred.

In all these situations, the aim is to assess whether the change or the status is statistically significant and to require that future laws, regulations and guidance assert the requirement to assess and report statistical significance.

4 Activities

4.1 Estimation of summary statistics[2)]

An estimate of a summary statistic depends on the values of water quality in the small volumes of water that happen to be captured by sampling and whether these values are measured accurately. The estimate, due to the overall uncertainty, is almost certain to differ from the true value of the summary statistic — the value that would be obtained if it were possible to achieve continuous error-free monitoring over the entire period for which the summary statistic applies.

Uncertainty can be managed by calculating confidence limits. Confidence limits define the range within which the true value of the estimate of the summary statistic is expected to lie. In the example in Table 1, the estimate of the mean from eight samples is 101 mg/l and there is a pair of 95 % confidence limits, 46 mg/l and 156 mg/l. There is 95 % confidence that the true mean exceeds the lower 95 % confidence limit of 46 mg/l and 95 % confidence that the true mean is less than the upper 95 % confidence limit of 156 mg/l. Overall there is 90 % confidence that the true value of the mean falls in the range between 46 mg/l and 156 mg/l[3)].

This range in the estimate of the mean represents large errors but these errors are seldom estimated or used to help take decisions based on the data. Also this discussion is for normally distributed random error. Such assumptions should be stated. Random error might not be normally distributed; it could be non-random and subject to mistakes and blunders. As a rule, the effect of these will be to increase the scale of the error. Errors should always be estimated even if this is done by making an assumption that they follow a normal distribution.

NOTE 1 The mean is used in this example because this summary statistic is commonly required by legislation. In other cases, there may grounds and opportunity to use other statistics like the median, e.g. to explain differences between large and small samples. The median is useful for data sets affected by outliers, and confidence limits can be calculated for the median.

---

2) Some documents use the concept of "sampling target". The sampling target could be the annual water quality, and a mean value over 1 year, or a 95-percentile over 1 year, is what is estimated.

3) This range is sometimes called the 90 % confidence interval, calculated from: $\overline{X} \pm t\sigma_{\overline{X}}$, where $\overline{X}$ is the mean; t is derived from the t-distribution with $n - 1$ degrees of freedom, used instead of the normal standard deviation for low rates of sampling (ISO/IEC Guide 99:1993[5]); and $\sigma_{\overline{X}}$ is the "standard error" derived from the standard deviation divided by the square root of the number of samples, $\sigma/\sqrt{n}$.

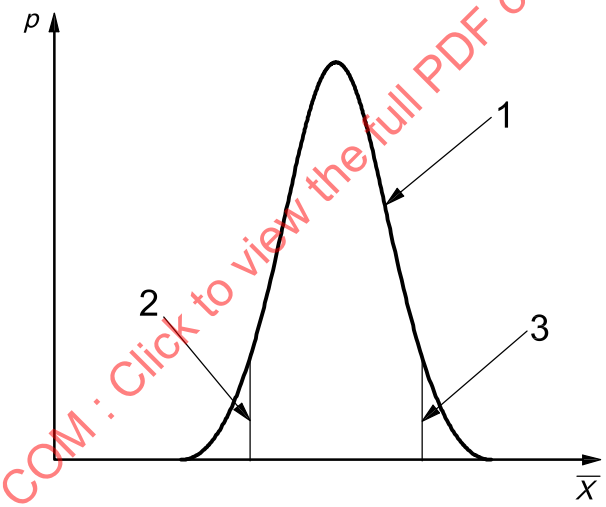
NOTE 2 Assumptions should always be stated. In this case, a normal distribution of errors is assumed.

NOTE 3 There are differences between large and small samples, i.e. use of the t-statistic versus the standard normal deviate, z.

Figure 2 illustrates the range of uncertainty. This range is an estimate of the distribution of errors in the estimate of the mean. The confidence limits are shown as points that mark 5 % and 95 % of the area of this distribution.

Table 1 — Example of confidence limits for the mean

Parameter	Value
Estimate of the mean	101 mg/l
Standard deviation	82 mg/l
No. samples	8
Lower confidence limit	46 mg/l
Upper confidence limit	156 mg/l



Key

p probability
$\overline{X}$ value of the mean
1 distribution of errors in the estimate of the mean
2 lower confidence limit
3 upper confidence limit

Figure 2 — Confidence limits on the mean

The gap between the confidence limits widens as the sampling rate is decreased (and would vanish for a continuous error-free monitor). It is also larger for estimates from the same number of samples, of more extreme summary statistics such as the 95-percentile and 99-percentile. For a typical water pollutant, the confidence limits for a mean estimated from 12 samples are ±30 %. For an estimate of the 95-percentile, this range is −20 % to +80 %.

4.2 Thresholds for water quality and compliance

The impact of the overall uncertainty makes it vital for most applications that thresholds and similar controls are defined or used as means or percentiles and, not, for example, as absolute limits (see 5.3, which identifies some circumstances where absolute limits may be applicable). In other words, the decision to use sampling means that definitions of thresholds should be restricted to summary statistics that can be assessed properly by sampling. In this clause, the discussion is restricted to summary statistics that are means and percentiles (but see also Clause 7).

NOTE The use of a threshold that is expressed, for example, as an annual mean implies that the pollutant causes damage that builds up over time. The annual mean can also apply if the impact of the pollutant is associated with values higher than the mean, so long as the shape of statistical distribution can be expected to be fairly stable across the situations where the threshold is used, and if action to reduce the mean will also reduce the number or scale of peak events in a reasonably regular way. The extent to which this is a risk to water quality is often covered by the size of the safety factors built into the threshold in the first place — a consideration for topic a) of the introduction. Given these conditions, the use of a mean as a threshold has the advantage that it is generally efficient in terms of getting the smallest overall sampling uncertainties from a set of samples.

4.2.1 Thresholds expressed as the mean

If the threshold is defined as a mean (e.g. representative of a period of time, such as 1 or more years or a season), then it is a simple matter to estimate the summary statistic from a set of samples. This estimate can then be compared with the value of the mean that is set down as the threshold. If the value estimated from the samples is worse than the value of the threshold, the site under test can be said to have failed. If the estimate is better than this value, the site can be said to have passed. This type of assessment is called a face-value assessment. It takes no account of errors.

A difficulty with the face-value assessment is that any estimate of the mean depends on the values captured by sampling and subsequently measured by an instrument or in a laboratory. There is a risk, caused by the overall uncertainty, that a compliant site (one which met the mean threshold) might be reported as a failure purely because the set of samples happened by chance to capture a few high values. Similarly, a non-compliant site might evade detection if it happened by chance to hold mainly good quality samples.

This means that the overall uncertainty carries a risk of making bad decisions. This risk can be controlled by allowing for the range that the overall uncertainty places around the estimate of the mean. One way of doing this is to calculate confidence limits (see Table 1).

In Table 1, the face-value estimate of the mean is 101 mg/l. Around this there is a pair of confidence limits, 46 mg/l and 156 mg/l. These define a confidence interval. With a 90 % confidence interval, there is 95 % confidence that the true mean is less than the upper confidence limit (156 mg/l in Table 1). There is a chance of only 5 % that the true mean water quality is as high as this. Similarly, there is 95 % confidence that the true mean exceeds the lower confidence limit (46 mg/l in Table 1). There is a chance of only 5 % that the true mean water quality is as low as this.

To assess compliance, the confidence limits should be compared with the threshold (all compliance assessment tests in this part of ISO 5667 are one tail):

— if high values of water quality are bad and the upper confidence limit is less than the mean threshold, there is at least 95 % confidence that the threshold was met;

— if high values of water quality are bad and the lower confidence limit exceeds the mean threshold, there is at least 95 % confidence that the threshold was not met;

— where the mean threshold lies between the upper and lower confidence limits, it is not possible to state compliance or failure with at least 95 % confidence.

With the results shown in Table 1, the site would pass a threshold set with a mean of 160 mg/l and fail one with a mean of 40 mg/l. These decisions have at least 95 % confidence. Performance against a mean threshold of 100 mg/l is unresolved at 95 % confidence — the value of 100 mg/l lies between the confidence limits.

Note that for the example in Table 1 (which uses eight samples), 95 % confidence of failure of a mean threshold of 46 mg/l (i.e. when the mean threshold equals the lower 95 % confidence limit) is not demonstrated by this method until the face-value estimate of the mean exceeds 101 mg/l. This is more than twice the threshold. If this is unacceptable, one of the remedies is to increase the number of samples.

4.2.2 Thresholds expressed as a percentile

A threshold that is expressed as a percentile is perhaps preferred to the mean, e.g. for damage associated with higher concentrations than the average. It can also apply if the impact of the pollutant is associated with values higher than the 95-percentile. This can happen so long as the shape of statistical distribution can be expected to be fairly stable across the situations where the threshold is used, and if action to reduce the 95-percentile will also reduce the number of peak events in a reasonably regular way.

Similarly the use of a limit like the annual 95-percentile implies knowledge or assumption that the duration of individual events of high concentration is not important so long as the total exceedence is less than 5 %, though a threshold expressed as an annual 95-percentile can also apply where the distribution of the duration of events is expected to remain fairly stable across the situations where the threshold is used.

NOTE To repeat, the risk to water quality of such an assumption is often covered by the size of the safety factors built into the threshold in the first place — a consideration for topic a) of the Introduction.

Where none of this applies, other types of threshold can be used, though these will imply that compliance might need to be assessed by monitoring that is nearly continuous, and not by a small set of samples. And if failure of such a limit constitutes an immediate emergency rather than advanced warning of a possible problem, it implies a capacity for real time control of the source of damage.

4.3 Confidence of failure

In taking decisions, the response could vary from “report the failure” to “take legal action” to “spend a lot of money” to “rectify the problem regardless of cost”. The consequences of being wrong vary, and, in principle, each type of decision requires its own degree of confidence, i.e. its own accepted risk of being wrong. The more important the decision, the less the decision maker should allow the play of uncertainty and errors in sampling and measurement to lead to a wrong decision.

The confidence of failure is a single statistic that replaces the need to compute different confidence limits for each type of decision. It varies on a scale from 0 % to 100 % (see Table 2).

Table 2 looks again at the data in Table 1. It shows the confidence of failure for three different mean thresholds — 180 mg/l, 120 mg/l, and 30 mg/l. For the mean threshold of 30 mg/l, the confidence of failure is 98 %. This means that there is a risk of only 2 % that the site under test met the mean threshold of 30 mg/l, but it appeared to fail because the action of chance produced a set of high results. In this case, it is appropriate to take any action where it is acceptable to live with a risk of up to 2 % that such action is truly unnecessary.

This is illustrated in Figure 3. The confidence of failure is the area of the uncertainty in the estimate of the mean that is cut by the threshold.

NOTE The calculation of confidence of failure is best done by computer (Reference [13]). It involves calculating a confidence limit on the estimate of the mean that would be represented by the threshold. For example, in the last line of Table 2, the threshold of 30 mg/l is a 98 % confidence limit.

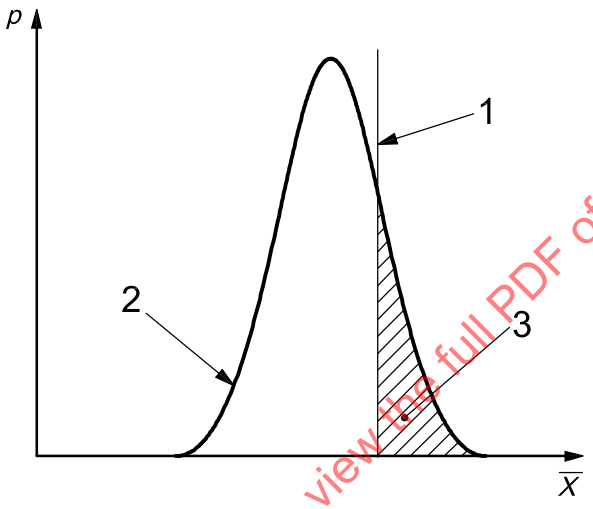
Table 2 looks at the risk that failure is wrongly reported (a “type I error”). The parallel exercise, though less common, is to look at the risk that success is claimed wrongly (a “type II error”). In Table 2, with the mean threshold of 180 mg/l, there is 1 % confidence that the threshold is failed. In other words there is 99 % confidence that it was met.

4.4 Methods for thresholds expressed as percentiles

One way of estimating percentiles from the results of sampling is to use an assumption about the statistical distribution from which the samples came, e.g. whether it is lognormal or normal. Such methods are called parametric methods, as distinct from non-parametric methods (that generally need to make no assumption about distribution).

Table 2 — Example of confidence of failure in water quality thresholds expressed as a mean

Parameter	Value
Estimate of the mean	101 mg/l
Standard deviation	82 mg/l
No. samples	8
Confidence of failure	**Percentage**
For a mean threshold of 180 mg/l	1
For a mean threshold of 120 mg/l	27
For a mean threshold of 30 mg/l	98



Key

p probability
$\overline{X}$ value of the mean
1 threshold
2 distribution of errors in the estimate of the mean
3 area showing confidence of failure

Figure 3 — Confidence of failure

Parametric methods depend on the fact, for example, that the 95-percentile for a normal distribution is 1,64 multiples of the standard deviation above the mean. An example in Table 3 gives an estimate of the 95-percentile of 250 mg/l for a mean and standard deviation of 101 mg/l and 82 mg/l respectively. In this example a lognormal distribution is assumed and the mean and standard deviation are converted to the log domain using the method of moments. The lower and upper 95 % confidence limits are 160 mg/l and 760 mg/l (calculated in this case using the properties of the shifted t-distribution — see Annex A).

NOTE The method of moments provides equations that convert the mean and standard deviation into estimates of the mean and standard deviation for the logarithms of the data without the need to take logarithms of the sample results themselves (see Annex A).

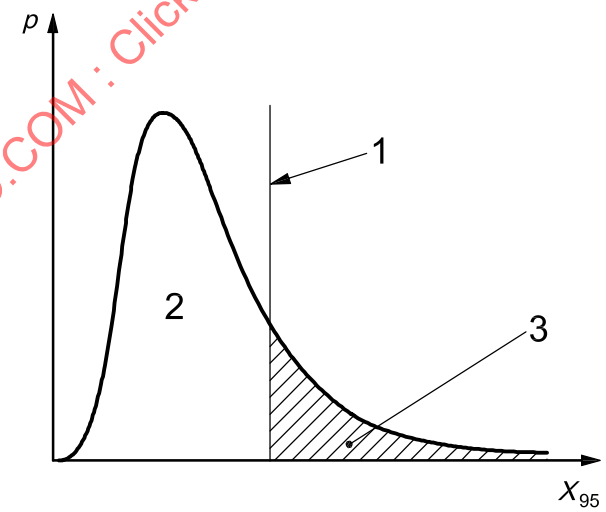
In Table 3, the confidence of failure for a 95-percentile threshold of 800 mg/l is 4 %. This states that there is 96 % confidence that the 95-percentile limit of 800 mg/l was met. There is a risk of only 4 % that the site truly met the threshold of 800 mg/l, but appeared to fail because the set of samples contained, through chance, an unexpected number of high values. For a limit of 300 mg/l, the confidence of failure is 40 %.

This is illustrated in Figure 4. The confidence of failure is the area of the uncertainty in the estimate of the 95-percentile that is cut by the threshold.

It is not the intention in the example of Table 3 to advocate the assumption of a lognormal distribution in circumstances where this is not appropriate, or to promote the use of the particular methods of calculating percentiles and confidence limits that are used for this example. But it is informative to do these sorts of calculations, despite the assumptions, to determine the scale of the error. This is better than ignoring the error, or pretending it is zero. More accurate methods of estimating the error can be used if this proves necessary. What is being advocated here is that it is wrong to act as if the true 95-percentile in the example in Table 3 were 250 mg/l and to act in ignorance of the fact that the range on this was something like 160 mg/l to 760 mg/l, or bigger.

Table 3 — Example of confidence of failure for water quality thresholds expressed as a percentile

Parameter	Value
Estimate of the mean	101 mg/l
Standard deviation	82 mg/l
No. samples	8
Estimate of 95-percentile	250 mg/l
Lower confidence limit on the estimate of 95-percentile	160 mg/l
Upper confidence limit on the estimate of the 95-percentile	760 mg/l
Confidence of failure	**Percentage**
For a 95-percentile threshold of 800 mg/l	4
For a 95-percentile threshold of 300 mg/l	40
For a 95-percentile threshold of 150 mg/l	96



Key

p probability
X_{95} value of the 95-percentile
1 threshold
2 distribution of errors in the estimate of the 95-percentile
3 area showing confidence of failure

Figure 4 — Confidence of failure for thresholds expressed as a percentile

It could be important in the context of the decisions made using a result of such data to use different statistical techniques. It might also be that the data are unrepresentative in time or space, or that there were mistakes in the mechanics of taking and handling the samples. Some of the data can be affected by limitations in analytical technique and expressed as “less than” some detection limit. There could be underlying trends. Many of these factors will mean that the overall error is even bigger than that suggested by the range from 160 mg/l to 760 mg/l.

Note that for the example in Table 3, with eight samples, 95 % confidence of failure of a 95-percentile threshold of 160 mg/l (i.e. when the 95-percentile threshold equals the lower 95 % confidence limit) is not demonstrated by this method until the face-value estimate of the 95-percentile exceeds 250 mg/l. A 250 mg/l concentration is 56 % bigger than the threshold. If this is unacceptable, one of the remedies is to increase the number of samples. For 26 samples, the 56 % is reduced to 34 %.

The assumption of lognormality is appropriate if data can be assumed to be roughly compatible with the lognormal distribution. The assumption extracts extra information from the calculation and so can boost the precision of estimates of percentiles, when compared with non-parametric methods of estimating percentiles.

4.5 Non-parametric methods

There are instances where parametric methods cause difficulties. It has been noted in 4.4 that it might sometimes be wrong to assume, for example, a lognormal distribution.

Non-parametric methods for the estimation of percentiles are based on ranking the sample results from smallest to largest. An estimate of the 95-percentile is given as the value that is approximately 95 % of the way along this ranked list, interpolating where this point falls between a pair of samples.

Since assumption (or information) is excluded, the estimates of confidence limits from this type of non-parametric method tend to be wider than those from the corresponding parametric methods.

The non-parametric methods can help in applications that involve legal actions where there is a need to avoid assumptions that might be contested, e.g. whether the sample results follow a lognormal distribution (or any other distribution).

NOTE Estimates from non-parametric methods might be less risky than a false assumption that the data are from a particular distribution. A parametric method might be better and more powerful if departures from such an assumption are unimportant. Whether a parametric or a non-parametric approach is better depends on the situation. An approach that could be considered is to use a parametric method that has the most flexible assumptions and to confirm this by applying a robust non-parametric method to representative examples. For both parametric and non-parametric it is important that the assumptions are identified.

In assessing compliance with a 95-percentile threshold, the estimate of the 95-percentile from a non-parametric can be compared with the threshold. As before, it is important to take account of the uncertainty in the estimate of the 95-percentile by calculating confidence limits or the confidence of failure.

When assessing compliance with a threshold expressed as a percentile, an alternative and simpler way of using a non-parametric approach is to count the number of failed samples, i.e. the number of sample results whose concentration exceeds the concentration in the percentile threshold. The proportion of failed samples is an estimate of the proportion of time spent in excess of the threshold. If more than 5 % of samples exceed the concentration in a 95-percentile threshold, it is tempting to say that the threshold was not met. However, this is a face-value assessment of the percentage of time spent outside the threshold, vulnerable to the errors expressed in the overall uncertainty.

This method of using data, counting failed samples, means that some of the information in the samples is not used, and this can be important. For example, there is no difference between a sample that only just exceeds a threshold and one that exceeds it grossly. Both are treated equally under this method, as no more than failed samples. Similarly, a sample that nearly fails the threshold is equivalent to one with a concentration of zero. Both are just compliant samples.

For 26 samples with one failed sample, the percentage of failed samples is $(1/26) \times 100$ or 3,85 %. This is an estimate of the true failure rate, i.e. the true time spent in failure. The value of 3,85 % is less than 5 %. This states, at face value, that a 95-percentile threshold has not been failed (whereas a 99-percentile limit would have been failed because 3,85 % is bigger than 1 %).

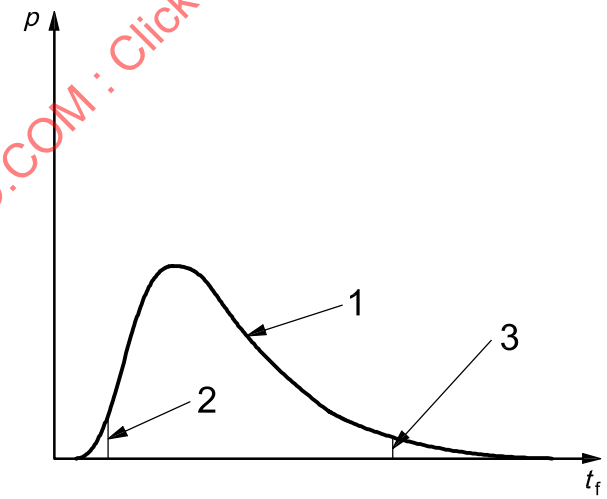
The third row of Table 4 shows that for 26 samples and one failed sample, the 95 % confidence limits about the value of 3,85 % are wide. They span 0,20 % and 17,0 % (in this particular case estimated from the properties of the binomial distribution — see Annex B). The 0,20 % and 17,0 % values define the lower and upper 95 % confidence limits on the estimate of the time spent in failure (the percentage of failed samples). There is 90 % confidence that the true failure rate is in this range.

The lower confidence limit shows that there is a risk of 5 % that a result as bad or worse than one failed sample in a set of 26 could have been produced from a site whose true failure rate was as low as 0,20 % of the time. Similarly, for the upper confidence limit there is a risk of only 5 % that a result as good or better than one failed sample in a set of 26 could have been produced from a site whose true failure rate was as bad as 17,0 % of the time.

Figure 5 illustrates this in terms of a curve showing the range of uncertainty, the distribution of errors, in the estimate of the time spent in failure. The confidence limits are shown as points that mark 5 % and 95 % of the area of the distribution. As before, a sampling process is used to estimate a summary statistic that in this case refers to the time spent in failure. Again the use of sampling introduces uncertainty.

Table 4 — Non-parametric method for percentiles

No. samples	No. failed samples	Percentage of failed samples %	True failure rate (90 % confidence interval) %
4	1	25,0	1,27 to 75,1
12	1	8,33	0,43 to 33,9
26	1	3,85	0,20 to 17,0
52	1	1,92	0,099 to 8,8
150	1	0,67	0,034 to 3,1



Key

p probability
t_f percentage time in failure
1 distribution of errors in the estimate of the time spent in failure
2 lower confidence limit
3 upper confidence limit

Figure 5 — Non-parametric method for percentiles

Table 4 shows that if there were 12 samples and one of the samples exceeded the threshold, there is 8,33 % of failed samples. This is the face-value estimate of the time spent in failure. Table 4 gives the corresponding lower and upper 95 % confidence limits as 0,43 % and 33,9 %.

Suppose the threshold was a 95-percentile. This means that the time spent in failure is not to exceed 5 %. As before, it is necessary to compare not only the face-value estimate of 8,33 with the allowance of 5 %, but to do the same with the lower confidence limit. If this exceeds 5 % there is at least 95 % confidence that the site has truly failed the 95-percentile threshold. In the case of 12 samples and 1 failed sample, the lower confidence limit is only 0,43 %. This is much smaller that 5 % and so the failure of the 95-percentile threshold suggested at face-value by the simple per cent of failed samples (8,33 %), is not significant at 95 % confidence.

The above deals with the assessment of failure. To be sure of a pass the upper confidence limit should be less than the 5 % set by the 95-percentile threshold. In any other position, where the value of 5 % lies between the two confidence limits, compliance is unresolved at 95 % confidence.

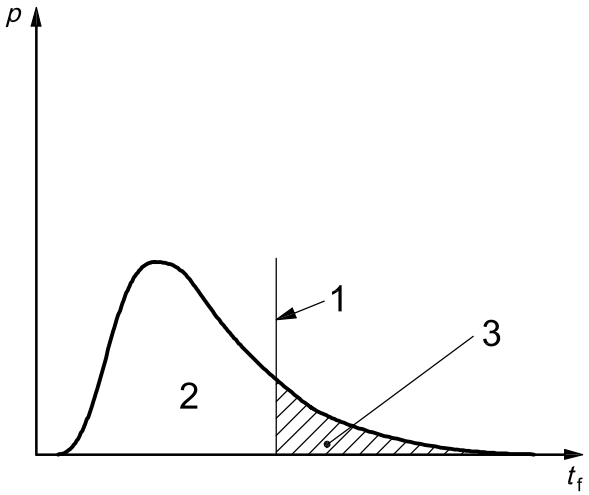
Table 5 parallels the contents of Table 4 but gives, for particular outcomes from sampling, the confidence of failure of a threshold expressed as a 95-percentile.

The second row of Table 5 (for calculation details, see Annex B) shows the case for a set of 12 samples in which one sample exceeded the threshold concentration set as the 95-percentile. Table 5 shows that there is 54 % confidence that this outcome, 12 samples of which one failed, indicates that the 95-percentile threshold has been failed, i.e. that the concentration defined as part of the 95-percentile threshold has been exceeded for more than 5 % of the time. In this case, it is appropriate to take decisions as a consequence of this set of samples, so long as it is acceptable that the risk is 46 % that the action is unnecessary. This might rule out expensive or irreversible decisions.

The third row of Table 5 shows the case for a set of 26 samples in which one sample exceeded the threshold concentration set as the 95-percentile. There is 26 % confidence that the 95-percentile threshold has been failed. This particular case is illustrated in Figure 6. The confidence of failure is the area of the uncertainty in the estimate of the time spent in failure that is cut by the threshold (5 %).

Table 5 — Confidence of failure of a water quality threshold expressed as a 95-percentile

No. samples	No. failed samples	Confidence of failure %
4	1	81
12	1	54
26	1	26
52	1	7
150	1	0,05



Key

p probability
t_f percentage time in failure
1 threshold (allows 5 % of failure)
2 distribution of errors in the estimate of the mean
3 area showing confidence of failure of 26 %

Figure 6 — Confidence of failure of a threshold expressed as a 95-percentile

4.6 Look-up tables

The use of the upper confidence limit, or the confidence of failure, to determine which sets of samples show failure that is statistically significant was discussed in 4.5. This process can be simplified by allowing more failed samples than, for example, the 5 % suggested by a 95-percentile threshold.

In this, the permitted number of failed samples is increased to a point where there is at least 95 % confidence that the site has failed for the percentage of time allowed by the percentile threshold, i.e. exceeds 5 % for the 95-percentile. This gives at most a risk of 1/20 that a site is wrongly declared as a failure. Table 6 contains values for the 95-percentile threshold and 95 % confidence of failure

Table 6 — Look-up table for the 95-percentile

No. samples	Minimum No. failed samples for at least 95 % confidence of failure
4 to 7	2
8 to 16	3
17 to 28	4
29 to 40	5
41 to 53	6
54 to 67	7

A table similar to Table 6 forms part of the permits sanctioning discharges from sewage treatment plants under Council Directive 91/271/EEC[7], which defines “failure” as a state where there is 95 % confidence that the 95-percentile threshold was failed.

Note that for eight samples, 95 % confidence of failure for 5 % of the time is not demonstrated by this non-parametric method unless at least three of the eight samples (or 37,5 %) have failed. If this is unacceptable, one of the remedies is to increase the number of samples.

Look-up tables can be set up for any combination of percentile and required confidence of failure. Table 7 gives values for the 99,5-percentile and 95 % confidence of failure and Table 8 gives a version for the 95-percentile and 99,5 % confidence of failure (these calculations exploit properties of the binomial distribution — see Annex B).

Table 7 — Look-up table for a 99,5-percentile threshold

No. samples	Minimum No. failed samples for at least 95 % confidence of failure
1 to 10	1
11 to 71	2

Table 8 — Look-up table for a 95-percentile threshold

No. samples	Minimum No. failed samples for at least 99,5 % confidence of failure
3 to 7	3
8 to 14	4
15 to 23	5

Just as extra failures are allowed in order to give proof of failure, so fewer failures than the 5 % associated with the 95-percentile threshold is a condition of demonstrating proof of success. In the design of rules in awarding prizes for proven compliance, a different look-up table is needed. This type of look-up table is designed to control the risk of stating wrongly that a site has truly failed and reported wrongly as compliant because of the overall uncertainty.

There is a limitation in using this type of look-up table to show high confidence of compliance with thresholds such as the 95-percentile. This is because, at low rates of sampling, reaching the required level of confidence can appear to require fewer than zero failed samples. Confirming at least 95 % confidence that a threshold concentration is met for 95 % of the time cannot be done with less than 57 samples.

NOTE The use of fewer than 57 samples can make it necessary to use parametric methods to prove compliance with percentile thresholds at high confidence.

5 Definition of thresholds

5.1 General

Water quality can be assessed by the use of thresholds. The results from samples are compared with these in order to assess compliance. This clause looks at how thresholds should be defined in order to avoid poor decisions about compliance.

5.2 Ideal thresholds

Difficulties can be avoided if thresholds are defined or treated as ideal thresholds, i.e. values that address five criteria, the first three of which are:

a) a level, e.g. a concentration of 10 µg/l;

b) a summary statistic, e.g. how often the level can be allowed to be exceeded, e.g. 1 % of the time;

NOTE This is the annual 99-percentile. Thresholds might also be expressed as other percentiles and averages for a particular period of time, e.g. a month.

c) the period of time over which this statistic applies, e.g. a calendar year.

These three criteria are of primary importance and set the threshold. A fourth is relevant when deciding the action to improve water quality. When the action is finished, the question arises: what residual risk of failure is acceptable in the long run, for example, as a consequence of rare patterns in the weather? The fourth point in the ideal threshold covers this:

d) the definition of the design risk, i.e. the proportion of time periods for which failure to meet the criteria in a), b), and c) above is accepted, (e.g. 1 year in 20).

In other words, using the numbers introduced so far, it is acceptable that an annual estimate of the 99-percentile exceeds 10 µg/l in 1 year in 20.

A fifth criterion deals with the actual assessment of compliance, i.e. from the samples taken in a particular calendar year:

e) the statistical confidence with which non-compliance is to be demonstrated before failure is reported or particular action taken.

A compromise is necessary between d) and e). Perhaps d) is not as important as a), b), and c), being better regarded as the long-term outcome given even continuous error-free monitoring. It relates to the acceptability of the physical consequences of truly failing the threshold. Compliance is dealt with in e). Failure might be defined as the case where the monitoring or sampling shows at least 95 % confidence that the failure is true and not attributable to the effect of chance and uncertainty in the sampling process.

In other words, again using the numbers introduced so far, it is acceptable that an annual estimate of the upper 95 % confidence limit exceeds 10 µg/l in 1 year in 20.

In an ideal threshold, all five criteria are defined explicitly. If any is left undefined, its value would take an arbitrary, unknown value that could vary from decision to decision as the threshold was used. It is unacceptable to state that the limit is 10 µg/l, whilst allowing any or several of the other four criteria to vary in an unknown manner for each decision.

Examples of ideal thresholds include the following.

— Over a long period of time, e.g. 20 years, the 95-percentile value of the concentration should be less than 200 µg/l in 19 summers out of 20 and failure will be declared when monitoring shows non-compliance with 95 % confidence.

— Over a long period of time, e.g. 20 years, the mean value of the concentration should be less than 0,6 mg/l in 5 years out of 10 and failure will be declared when monitoring shows non-compliance with 95 % confidence.

5.3 Absolute limits

The purpose of a threshold is compromised if it is defined in a way that ignores the fact that compliance will be checked by sampling. One type of threshold that runs this risk is the maximum value (or absolute limit). This type of threshold is popular because it is easy to understand and use, especially in legal actions.

These benefits should be set against the extra errors that arise when maxima are assessed against data collected by sampling. These errors can lead to faulty assessments of performance and so to wrong decisions, e.g. on legal action, on investment to improve quality that does not, in reality, require improvement or on failure to invest where improvement is truly necessary.

In particular, great care should be taken over absolute limits and over the number of data used to assess compliance. This is because a relatively small number of samples is taken from a relatively wide range of time

and material. This leads to a strong risk that there will be high concentrations (and exceedences) outside the instances captured in the samples. The fact that few failures are seen can encourage regulators to position the thresholds in absolute limits at values that seldom elicit failure under, say, a monthly sampling regime.

The problem is that:

— increasing sampling will lead to more failed samples; and

— a report that the threshold has been failed is almost guaranteed under continuous monitoring or very frequent sampling.

To illustrate, consider a site that exceeds a threshold for 1 % of the time. Such a site will always be reported as a failure of the absolute limit if assessed using a continuous error-free monitoring. Table 9 shows that if assessed by sampling, this failure will escape detection with a probability that depends strongly on the number of samples.

EXAMPLE These calculations are based on the probability of no failures in a set of n samples. This is $(1 - P)^n$, where P is the probability of a compliant sample. Thus for 12 samples in Table 8, this becomes $(1 - 0{,}99)^{12} = 10^{-24}$.

With four samples there is 4 % probability that at least one of them exceeds the absolute limit. This rises to 39 % for 52 samples.

In this situation, the illusion of improved performance can be manufactured by taking fewer samples. In the meantime, the "true quality" might have deteriorated.

Table 9 — Effect of sampling rate on reported compliance

No. samples	Probability of reporting failure %
4	4
12	11
52	41

As discussed below, absolute limits monitored solely by sampling are not true absolute limits at all. This is because of the mathematical implication that failure is permitted at times when the samples are not taken, i.e. failure is tolerated for a proportion of the time. Such absolute limits are, in truth, percentiles.

When seeking a solution to the problems caused by an absolute limit, it is necessary to:

— translate it into a specific percentile in the permits and regulations; or, if this is not possible,

— treat it as a percentile when assessing compliance.

As an ideal threshold, the absolute limit has the required clarity for the first item, which might be a value, e.g. a concentration of 10 µg/l. However, for the second item, the summary statistic is ambiguous. The absolute limit requires compliance by 100 % of samples in a year.

This has two meanings. The first meaning is that the limit is a 100-percentile, a value that should really be met for 100 % of a year. It has been discussed above that this is illogical if sampling alone is used to assess compliance. The use of only 12 samples leaves a lot of time where failure could have occurred and might not have been observed.

This second option, treating the absolute limits as, for example, 99,5-percentiles, is attractive in cases where the limit has been set so strictly that occasional failed samples are likely, but where an infrequent failure is of low concern. This option controls the problem, illustrated in Table 10, that the percentile (and the severity of the limit or threshold) changes with the sampling rate. It also avoids the untidy and uncomfortable alternative of inventing rules by which operators and regulators discount certain failed samples.

Table 10 — Effect of sampling rate on the severity of an absolute limit

No. samples	No. failed samples	Equivalent percentile	Confidence of failure %
4	1	84,1	50
12	1	94,4	50
52	1	98,7	50

The second possible meaning relies on the fact that an absolute limit coupled with a sampling rate actually defines candidate pairings of a percentile and a level of proof for declaring failure. For example, consider 12 samples and a rule where none of these is permitted to exceed the threshold. This outcome is exactly the same as a threshold that has been set, e.g. as a 95-percentile concentration that requires that sampling demonstrate 50 % confidence of failure before a site is declared to have failed. The same rule, 12 samples and no failures, is also equivalent to a 99,5-percentile concentration with a 95 % level of proof [4]. Any number of other pairings of percentile and level of proof is possible.

An absolute limit of, say, 10 µg/l, is equivalent to a 75-percentile if assessed from 4 samples, but a 98-percentile if checked against 52 samples. This change in percentile with sampling rate is equivalent, typically, to a move from 10 µg/l to 30 µg/l in the applied threshold. This should be considered an arbitrary and unfair change in severity.

Similarly, for a fixed percentile, the severity of the threshold is increased in terms of the degree of proof required to produce a report of failure (see Table 11).

Table 11 — Effect of sampling rate on the severity of an absolute limit

No. samples	No. failed samples	Equivalent percentile	Confidence of failure %
4	1	95	81
12	1	95	54
52	1	95	7

These problems are controlled if the limit is set, as for an ideal threshold, to some particular combination of percentile and level of proof, i.e. the 99,5-percentile with a level of proof of 95 %.

However, it could be the case that the limit can never be failed because it is known to cause immediate damage. In this case, it can be best to move the threshold to a lower concentration, e.g. 4 µg/l as a 99-percentile instead of 10 µg/l as a “100-percentile”, perhaps retaining the original 10 µg/l as an additional control. In this case, compliance with the 99-percentile of 4 µg/l might be taken as acceptable confidence of not getting an actual value of 10 µg/l.

If the absolute limit is to be used on its own there is an implication that continuous accurate monitoring is required and that there are real-time controls that can respond to risk and prevent exceedence.

Also, it could be that an absolute limit has been set on the understanding that more information than sampling is used to judge performance. Such extra information might include the fact that high concentrations cannot occur for the type of process that is being monitored, or because other records kept by plant operators demonstrate that treatment plants are working well. There is a risk, however, that the sample results are used outside of these contexts.

4) These conclusions follow from the properties of the binomial distribution. One failed sample in a set of 12 gives a probability of greater than 95 % that the threshold was failed for at least 0,5 % of the time covered by the samples.

In adopting e.g. the 99,5-percentile instead of the “100-percentile”, the regulator acknowledges that it is impossible to demonstrate from sampling alone that a limit was met for 100 % of the time. This method, declaring the percentile point and level of proof, allows the taking of more or fewer samples, whilst retaining the same severity of threshold and level of proof.

The absolute limit has the advantage that lawyers and the public easily understand it. An absolute limit might be retained in law, but with a consideration that it is used within a declared policy that the basis for taking a decision, i.e. to prosecute for non-compliance, treats the absolute limit, for example, as a 99,5-percentile requiring 95 % confidence of failure.

5.4 Percentage of failed samples

If the concentration in the sample exceeds a certain threshold, then the sample can be said to fail. Some thresholds are expressed as the maximum percentage of failed samples in a set of samples taken over a period of time.

Such an approach betrays a lack of appreciation of the difference between a statistical population and a set of samples. A statistical population is the distribution of all the values that actually occur over a period of time, e.g. in 1 year. For 1 year it can be thought of as approximating to the set of error-free results of chemical analysis taken from totally representative samples taken for each of the 31 536 000 s in the year. In contrast, there might be only 12 samples. The results of such samples are used to make an estimate for the population, such as its annual mean.

As demonstrated by this part of ISO 5667, thresholds should be expressed as a function of the population, e.g. that the concentration should be below the threshold for at least 95 % of the time. Where they are expressed as thresholds to be met by, say, 95 % of samples, this should be done within a suitably adaptive framework, which acknowledges that the percentage of failed samples is only an estimate of the time spent in failure. As discussed throughout this part of ISO 5667, such estimates are subject to large uncertainties.

If a threshold is defined as a limit to be met by at least 95 % of samples and 8 % of samples fail, this value, 8 %, will mean that the site is declared to have failed because 8 % is greater than 5 %, the allowed percentage of failed samples.

Although the 8 % of failed samples exceeds 5 %, the conclusion that the time spent in failure exceeds 5 %, is certain only if sampling is continuous, representative and accurate. There might be only 12 samples in a year, and purely by chance, the true failure rate might have been less than 5 %, but a set of samples with high concentrations might have been collected. In this case the site under test would be wrongly condemned because of the overall uncertainty (and the decision to take only 12 samples).

Thresholds defined as having to be met by a percentage of samples should be defined and treated as the corresponding percentiles — a concentration required to be met by 95 % of samples should be treated as a 95-percentile. Confidence limits, or the confidence of failure, should then be calculated and action taken according to the accepted risk of acting unnecessarily. This is discussed in Clause 4 for Tables 1, 2 and 3.

5.5 Calculating limits for effluent discharges

Data collected by sampling are also used to set the limits needed in permits in order to meet thresholds. Table 12 shows the results of the calculation (by Monte-Carlo simulation, see Council Directive 91/271/EEC[7]) of the mean and 95-percentile of discharge quality needed in order to meet a 90-percentile limit of 1,3 mg/l in a river.

These calculations ensure that the permit conditions are justified in terms of being necessary to meet the environmental requirement, and that they go no further than this. The need to do these calculations reinforces the requirement for the ideal thresholds discussed in 5.2 and the need to define thresholds as means or percentiles, although the actual calculations for such permit conditions lie outside the scope of this part of ISO 5667.

Environmental problems might be due to short term events. For example, a river might be fully saturated with oxygen for more than 95 % of the time but a few minutes of total deprivation from oxygen will kill the fish. The use of a 5-percentile threshold in this case works only to the extent that the extreme events are correlated with

the 5-percentile (4.2) in the context of the safety factors built into the 5-percentile threshold. As discussed in 4.2 and 5.2, there can be cases where this does not apply, though if this risk is to be lived with and managed by thresholds, there is an implication that compliance cannot be assessed by sampling, but requires some form of continuous assessment (and, perhaps, real-time control).

Table 12 — Calculating limits for a discharge into a river with a downstream threshold set as a 90-percentile

Input data	
Mean river flow upstream of discharge	325,00 Ml/d
5-Percentile of river flow	40,00 Ml/d
Mean upstream quality	0,08 mg/l
Standard deviation	0,07 mg/l
Mean flow of discharge	59,00 Ml/d
Standard deviation	19,00 Ml/d
Present mean quality of discharge	7,20 mg/l
Standard deviation	4,10 mg/l
Results	
Mean river quality downstream of discharge	0,63 mg/l
90-Percentile river quality	1,30 mg/l
River quality threshold (90-percentile)	1,30 mg/l
Required mean quality in discharge	2,51 mg/l
Required 95-percentile discharge quality	5,22 mg/l

6 Declaring that a substance has been detected

A variation on the issue of absolute limits lies in answering questions such as, “was the substance detected by chemical analysis?” This should be answered by calculating, for example, whether there is at least 95 % confidence that the substance is present at or above an agreed limit of detection, for at least 10 % of the time. This can be done by using Table 13.

NOTE 1 Table 13 is a variation on Table 6, but for a limit that can be exceeded for 10 % of the time — a 90-percentile. See Annex B for the calculation method.

Table 13 — Defining “detected”

No. samples	Minimum No. samples at or above the limit of detection
2 to 3	2
4 to 8	3
9 to 14	4
15 to 20	5

Table 13 means, for example, that if 12 samples are taken and four of these exceed the limit of detection, there is at least 95 % confidence that the limit of detection was exceeded for 10 % of the time.

A parallel process might be to ask how many sample results below a detection limit are needed before it can be concluded that a substance is not present. Table 14 shows rules for showing with at least 95 % confidence that a substance is not present for more than a specific percentage of the time (see Annex B for the calculation method). For example, Table 14 means that absence is demonstrated for 50 % of the time, at

95 % confidence, if five samples are taken and all of them are below the limit of detection. To demonstrate absence for 99 % of the time at 95 % confidence, 298 samples must be taken and all must be recorded as below the limit of detection.

Table 14 — Confirming "absence"

To demonstrate absence for the following percentages of time	Minimum No. samples taken and all shown to be below the limit of detection
50	5
80	14
90	29
95	59
99	298

NOTE 2 It is a common misunderstanding that when laboratories flag data as below the detection limit that they are guaranteeing values to be lower than the specified concentration. In fact what laboratories may be reporting is that the uncertainty is so large at this concentration that a quantitative value cannot be reported. Thus "< 5 mg/l" is a flag that uncertainty is too large for reporting by the significant figures convention, but is not a guarantee that the result falls below 5 mg/l. Tables 13 and 14 need to be interpreted in this sense, if applied to such data.

NOTE 3 The use of Tables 13 and 14 implies that all the analyses were done by methods having consistent limits of detection. Otherwise the results for "detection" and "absence" cannot be interpreted with reference to a particular limit of detection.

7 Detecting change

Sometimes sampling is used to demonstrate that water quality has improved or not deteriorated. If the estimate of the mean was 20 mg/l in 2003 and 25 mg/l in 2004, this looks like an increase in mean concentration of 25 %.

As before, this conclusion is affected by the overall uncertainty. There is a need to calculate the statistical confidence that the recorded difference is significant. A test can be used to calculate the significance of an apparent difference in the mean. Table 15 gives an example.

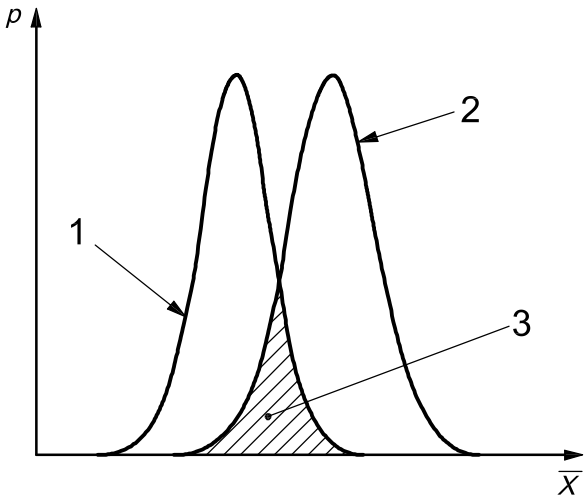
NOTE 1 The statistical test used for Table 15 is standard, based on the t-distribution. It assumes that the two distributions are normal and that each value is sampled independently from the others. There are many other methods; however, their discussion lies outside the scope of this part of ISO 5667 (see ISO/IEC Guide 99:1993[5]). It is important to estimate the statistical confidence of a difference, and not just to act as if there were no uncertainty, even if this involves the use of approximate techniques.

In Table 15, an apparent difference in the mean of 25 % is confirmed as significant at a level of confidence of 97 %. This indicates there is a chance of only 3 % that a difference as large as this could arise by chance.

This is illustrated in Figure 7. The confidence that no true change in mean has occurred is the area of overlap between the distributions of uncertainty for each estimate.

Table 15 — Assessment of change in mean

Year	Mean	Standard deviation	No. samples	Confidence of change %
2003	20	10	25	97
2004	25	9	33	



Key

p probability
$\overline{X}$ value of the mean
1 distribution of uncertainty in first estimate of the mean
2 distribution of uncertainty in second estimate of the mean
3 confidence the two means differ

Figure 7 — Change in mean

Similarly, if the estimate of the 95-percentile was 39 in 2003 and 42 in 2004, this looks like an increase of 8 %. Again this is the face-value conclusion. As in Table 15, it is necessary to calculate the statistical confidence that the recorded difference is real, as in the following example of the output from a parametric calculation (see Table 16).

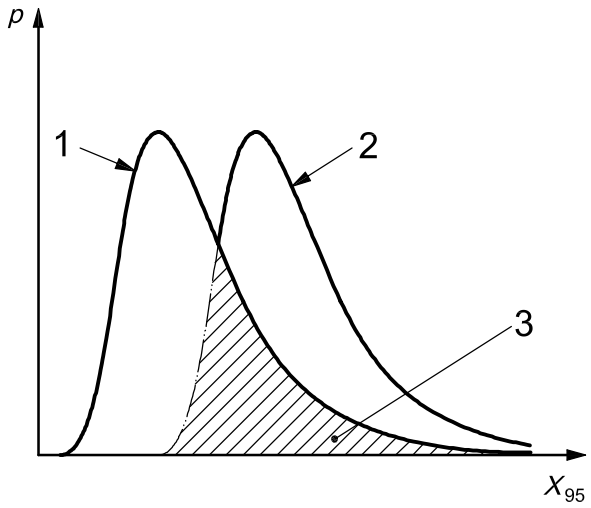
Table 16 states that there is a risk of 28 % that the increase from 39 to 42 is due to chance (and a 72 % probability that the change is a real one). This is illustrated in Figure 8. The confidence of that no true change in 95-percentile has occurred is the area of overlap between the distributions of uncertainty for each estimate.

NOTE 2 There are lots of tests for detecting change. The above example uses a test suitable for small sets of samples (a t-test). The aim of the test is to illustrate the need to consider the error using a technique that produces the sort of information in the tables. If the data are subject to things like seasonal cycles, more powerful techniques, including non-parametric methods, are available that are better able to pick out with more confidence than a t-test cases where the true water quality has changed by smaller amounts.

It is always useful to plot the data on graphs to look for effects like season cycles and to check whether the change is caused by one or two high values rather than a change across the board. This is important in assessing what can be done in response to the change.

Table 16 — Assessment of change in 95-percentile

Year	Estimate of 95-percentile	Standard deviation	Confidence of change %
2003	39	25	72
2004	42	33	



Key

p	probability
X_{95}	value of the 95-percentile
1	distribution of uncertainty in first estimate of the 95-percentile
2	distribution of uncertainty in second estimate of the 95-percentile
3	area showing confidence the two percentiles do not differ

Figure 8 — Change in 95-percentile

Non-parametric methods can be used to tackle the same issue of checking for change. In one method, the first step is to look at the uncertainties in the estimate of a limit, L, which might be an estimate of the 95-percentile for 2003, i.e. the concentration exceeded for 5 % of the time. This estimate might be based on n_{2003}, the number of samples used to estimate L, and E_{2003}, the number of exceedences of L in 2003.

Suppose in 2004, n_{2004} samples were taken and E_{2004} exceedences of L are observed. A test (e.g. using Fisher's exact test for 2×2 contingency tables) can be carried out to compare the proportion of exceedences in 2003 (E_{2003}/n_{2003}) with the proportion in 2004 (E_{2004}/n_{2004}). Table 17 illustrates the outcome.

Table 17 — A non-parametric assessment of deterioration

Year	No. samples	No. exceedences	Failed samples %	Confidence of change from 2003 to 2004 %
2003	40	3	7,5	93
2004	12	3	25,0	

In Table 17, the incidence of exceedence is apparently higher in 2004 (25,0 %) than in 2003 (7,5 %). At face value this is a deterioration. It turns out that there is a probability of 7 % that a discrepancy this big could have arisen by chance.

In summary, to demonstrate change, confidence of change should always be calculated. This helps distinguish changes that can be ascribed to the overall uncertainty, from those that really have occurred.

NOTE 3 If the data are subject to features like seasonal cycles, more powerful non-parametric methods are available that can pick out with more confidence than the tests used for Table 16 and Table 17 cases where the true water quality has changed by smaller amounts.

8 Classification

8.1 General

Sometimes water quality is described in terms of a classification system. In such a system there might be sets of thresholds, for one or more pollutants. Table 18 illustrates a classification system for a single pollutant. The class limits might be summary statistics of water quality, such as the annual 95-percentile.

Table 18 — Example classification system for a single pollutant using 95-percentile class limits

Class	95-percentile class limits (mg/l)
1	$\leqslant 10$
2	> 10 to $\leqslant 20$
3	> 20 to $\leqslant 40$
4	> 40 to $\leqslant 80$
5	> 80

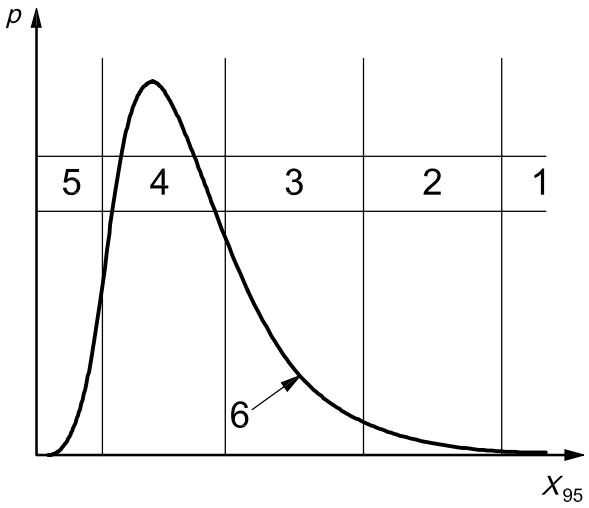
To assign the class, it might be that an estimate is made of the 95-percentile. If this value were 25 mg/l, it could be said that the site was in class 3 because 25 mg/l falls in the range from 20 mg/l to 40 mg/l (see Table 18). Again, this is a face-value assessment. The true 95-percentile might have differed from 25 mg/l, and the true class might have been class 2 or 4. This risk of a mistake is caused by the overall uncertainty.

As in the examples discussed above (e.g. in Tables 2, 3, 4, 13, 15, 16, and 17), it is important to estimate the statistical confidence that water quality exceeded the thresholds. In this case, this means the confidence that the estimate of the 95-percentile exceeded 40 mg/l, the confidence that the estimate was less than 20 mg/l, and the confidence that it lay between 20 mg/l and 40 mg/l. This might give the results listed in Table 19 and illustrated in Figure 9.

In Table 19, there is 56 % confidence that class 3 is the true class, 35 % confidence that the true class is class 4, 5 % that it is class 2 and even 4 % that it is class 5, i.e. two classes worse than the face-value class. The possibility of class 1 has zero confidence.

Table 19 — Example classification for a single pollutant using confidence of class

Confidence of class %				
Class 1	**Class 2**	**Class 3**	**Class 4**	**Class 5**
—	5	56	35	4



Key

p	probability
X_{95}	value of the 95-percentile
1	class 1
2	class 2
3	class 3
4	class 4
5	class 5
6	distribution of errors in estimate of the 95-percentile

Figure 9 — Classification for a single pollutant using confidence of class

In practice, the classification can be based on several pollutants and the "worst" result in terms of the classification might be used to define the class. Table 20 illustrates the case of four pollutants, A, B, C and D and the combination for these to give an overall class.

In Table 20, it is pollutant D that sets the face-value class, i.e. class 4. This has 68 % confidence. There is 4 % confidence of class 5 as a result of pollutant C. The residual confidence is 28 % [(100 – 68 – 4)]. In this particular case, this residual is assigned to the worst class of better quality than the face-value class — i.e. class 3.

Table 20 — Example classification system based on four pollutants

Pollutant	Confidence of class %				
	Class 1	Class 2	Class 3	Class 4	Class 5
Pollutant A	—	94	6	—	—
Pollutant B	—	23	76	1	—
Pollutant C	—	5	56	35	4
Pollutant D	32	—	—	68	—
Overall	—	—	28	68	4

The confidence of failure of a target class can be used to rank priorities — to manage the risk of action that later might turn out to have been unnecessary. For a target of class 1 and or class 2, Table 20 gives 100 % confidence of failure. For a target of class 3, the confidence of failure is 72 % [(68 + 4) %].

8.2 Confidence that class has changed

Suppose that the face-value class changed from class 4 in 1999 to class 3 in 2004. This looks like an improvement[5] (see Table 21).

Table 21 — Example confidence of class change

Confidence of class change %		
Class	**1999**	**2004**
1	—	—
2	—	40
3	20	60
4	70	—
5	10	—

The confidence of a change from class 4 to class 3 is the product of the values of the confidence of class for class 4 in 1999 and for class 3 in 2004, converted to a percentage by multiplying by 100, i.e. $0,7 \times 0,6 \times 100 = 42$ %.

All the possible combinations are given in Table 22. This shows 42 % confidence of the change from class 4 in 1999 to class 3 in 2004. It also shows 8 % confidence of a change from class 3 in 1999 to class 2 in 2004, 12 % confidence that the site stayed in class 3, and 4 % confidence of a change from class 5 to class 2. Finally, there is 6 % confidence in a move from class 5 to class 3.

Table 22 — Confidence of a change in class

		Class in 2004					**Confidence in 1999**
		1	2	3	4	5	%
Class in 1999	1	0	0	0	0	0	0
	2	0	0	0	0	0	0
	3	0	8	12	0	0	20
	4	0	28	42	0	0	70
	5	0	4	6	0	0	10
Confidence in 2004	%	0	40	60	0	0	

The sum of the numbers in the diagonal (dark-shaded) cells in Table 22 gives the overall confidence of no change in class. This is 12 %, i.e. in this case this is the same as the confidence that the site stayed in class 3. The entries are zero for no change from Class 1, 2, 4 or 5.

The diagonal sums of the adjacent lower (light-shaded) cells give the confidence of upgrades. There is 50 % confidence of an improvement by one class. This is made up of a 42 % confidence of a change from class 4 to class 3 and an 8 % confidence of a change from class 3 to class 2. Similarly there is 34 % confidence of an improvement by two classes.

5) This assumes low numbered classes are of good water quality.

The sum of the light-shaded cells in Table 22 shows the confidence of an improvement of one class or more to be 88 %. (Similarly the sum of the upper diagonals gives the confidence of an overall drop in class, which is zero.) Following this logic, the situation can be summarised as in Table 23. This shows 50 % confidence that quality improved by one class and 34 % confidence that the improvement was by two classes.

Table 23 — Example confidence of change in class

Change	Confidence (%)
Down 2 classes	0
Down 1 class	0
No change in class	12
Up 1 class	50
Up 2 classes	34
Up 3 classes	4
Up 4 classes	0

The data can also be presented as an accumulating sum, from the bottom, to give the numbers in Table 24. Tables 19 to 24 are real examples from the management of river water quality.

Table 24 — Confidence of a change in class

Change	Confidence
No downgrade	100
Up at least one class	88
Up at least 2 classes	38
Up at least 3 classes	4
Up at least 4 classes	0

It might be that over the period of assessment of change in class, that different methods and instruments were used. The effect of changes in the random errors in these will come through in the analysis. If in Table 22 the results for one period were based on fewer samples than the other period, or on less accurate methods of chemical analysis, this will come though as a wider spread of the confidence of class. This will give a reduced ability to establish significant changes in class between 1999 and 2004. On the other hand, the risk is controlled that expensive action to improve water quality, or complacency that all is well, is caused by errors in monitoring.

A more difficult issue occurs if the methods of 1999 or 2004 or both were biased or based on unrepresentative samples. This undermines the assessment of class and change in class. However, a lack of knowledge of all the errors is no excuse for failing to estimate the impact of the overall uncertainty.

Annex A
(informative)

Calculation of confidence limits

NOTE See ISO/TS 21748[3] and ISO/IEC Guide 98:1995[4].

As an example, a parametric method (method of moments, described in this annex) is used in 4.4 to estimate confidence limits around estimates of percentiles. The estimate of the mean of the logarithms of the data, M, and the estimate of the standard deviation of the logarithms of the data, S, are given by Equations (A.1) and (A.2):

$$M = \ln\left\{\frac{m}{\sqrt{1+(s^2/m^2)}}\right\} \tag{A.1}$$

$$S = \sqrt{\ln[1+(s^2/m^2)]} \tag{A.2}$$

where

m is the mean value of the measurements;

s is the standard deviation of the measurements.

The face-value estimate of the 95-percentile, $\widehat{X}_{95}$, is given by Equation (A.3):

$$\widehat{X}_{95} = \exp(M + zS) \tag{A.3}$$

where z is the standard normal deviate, which takes a value of 1,644 9 for 95 % confidence. To calculate confidence limits, z is replaced by t_0, a value which depends on the sampling rate. The two values of t_0 are given by Equation (A.4):

$$t_0 = \frac{\delta \pm \lambda\sqrt{1+(\delta^2/2\nu)-(\lambda^2/2\nu)}}{\sqrt{n}[1-(\lambda^2/2\nu)]} \tag{A.4}$$

where

n is the number of samples;

δ is given by

$$z\sqrt{n} \tag{A.5}$$

λ approximates to the normal standard deviate, i.e. 1,644 9;

ν is the number of degrees of freedom, in this case $n - 1$.

Take as an example a set of eight samples with mean 101 and standard deviation 82 (see Table 3):

$$M = \ln\left\{\frac{101}{\sqrt{1+(82^2/101^2)}}\right\} = 4{,}362\,0 \tag{A.6}$$