Environment Agency

using science to
create a better place

A comparison of Bayesian and Frequentist approaches for estimating WFD classification errors

Science Report: SC060044/SR1

SNIFFER
SCOTLAND & NORTHERN IRELAND
FORUM FOR ENVIRONMENTAL RESEARCH

ea/br/e/sci/v1

SCHO1208BPCQ-E-P

The Environment Agency is the leading public body protecting and improving the environment in England and Wales.

It's our job to make sure that air, land and water are looked after by everyone in today's society, so that tomorrow's generations inherit a cleaner, healthier world.

Our work includes tackling flooding and pollution incidents, reducing industry's impacts on the environment, cleaning up rivers, coastal waters and contaminated land, and improving wildlife habitats.

This report is the result of research commissioned and funded by the Environment Agency's Science Programme.

Published by:
Environment Agency, Rio House, Waterside Drive, Aztec West, Almondsbury, Bristol, BS32 4UD
Tel: 01454 624400 Fax: 01454 624409
www.environment-agency.gov.uk

ISBN: 978-1-84432-970-0

© Environment Agency February 2009

All rights reserved. This document may be reproduced with prior permission of the Environment Agency.

The views expressed in this document are not necessarily those of the Environment Agency.

This report is printed on Cyclus Print, a 100% recycled stock, which is 100% post consumer waste and is totally chlorine free. Water used is treated and in most cases returned to source in better condition than removed.

Further copies of this report are available from:
The Environment Agency's National Customer Contact Centre by emailing enquiries@environment-agency.gov.uk or by telephoning 08708 506506.

This report is the result of work jointly funded by the Environment Agency and the Scotland and Northern Ireland Forum for Environmental Research (Sniffer).

Author(s):
Julian Ellis, Robin Wyatt

Dissemination Status:
Publicly available

Keywords:
WFD, Bayesian approach, frequentist approach, spatial variability, temporal variability

Research Contractor:
WRc

Environment Agency's Project Manager:
Veronique Adriaenssens

Science Project Number:
SC060044

Product Code:
SCHO1208BPCQ-E-P

Science at the Environment Agency

Science underpins the work of the Environment Agency. It provides an up-to-date understanding of the world about us and helps us to develop monitoring tools and techniques to manage our environment as efficiently and effectively as possible.

The work of the Environment Agency's Science Department is a key ingredient in the partnership between research, policy and operations that enables the Environment Agency to protect and restore our environment.

The science programme focuses on five main areas of activity:

- **Setting the agenda**, by identifying where strategic science can inform our evidence-based policies, advisory and regulatory roles;
- **Funding science**, by supporting programmes, projects and people in response to long-term strategic needs, medium-term policy priorities and shorter-term operational requirements;
- **Managing science**, by ensuring that our programmes and projects are fit for purpose and executed according to international scientific standards;
- **Carrying out science**, by undertaking research – either by contracting it out to research organisations and consultancies or by doing it ourselves;
- **Delivering information, advice, tools and techniques**, by making appropriate products available to our policy and operations staff.

Steve Killeen

Steve Killeen

Head of Science

Executive Summary

Background / Need

The Water Framework Directive (WFD) states that all European surface waters should achieve good ecological status by 2015. Ecological status is an expression of the quality of the structure and functioning of biological elements associated with surface waters, classified in Annex V. Ecological status should be assessed using a reference condition approach and classification tools based on five biological elements (EU 2000).

To help achieve comparability of monitoring systems between the different Member States, each Member State is required to express the results from their monitoring systems as ecological quality ratios (EQRs) for the purposes of classification of ecological status. It is also stipulated that the precision and confidence achieved by the monitoring results must be quantified.

In developing the techniques required to implement this system, the Environment Agency and SNIFFER have collaborated on related R&D projects that are investigating sources of uncertainty in the application of the classification tools and the implications for the reliability of the classification schemes. To date, this work on quantifying the uncertainty in WFD water body assessment has mostly focused on a ‘classical’ or ‘frequentist’ statistical approach. This has led to the Confidence of Class (CofC) method, as described by Ellis & Adriaenssens (2006) and subsequently used by many of the WFD tool developers.

In a follow-on project entitled: ‘Uncertainty estimation for monitoring for each of the WFD biological classification tools – Further work on classification, uncertainty and variability aspects‘, the question was raised as to whether or not an alternative to the Confidence of Class approach could be the application of a Bayesian statistical approach. Accordingly it was decided to select a WFD tool for which there was a good set of historical data and to use this dataset to demonstrate the Bayesian approach. The data would also be analysed using the existing CofC approach, thus enabling the two sets of results to be contrasted.

Main objectives / Aims

A comparative analysis has been carried out for the DARES diatom tool using both a frequentist approach (as currently adopted by most of the WFD tools) and a Bayesian approach (as adopted by the Fisheries WFD tool).

Results

We have used the two statistical approaches to analyse illustrative data sets for each of seven ‘new’ sites. This has shown that, provided similar statistical assumptions are made in analysing the data for the 105 historical sites, there is very little practical difference between (a) the Confidence of Class values generated by the frequentist approach for the new sites and (b) the corresponding Probability of Class values arising from the Bayesian method.

More generally, the exercise has demonstrated the ease with which the Bayesian approach (allied to the WinBUGS software) can be adapted to incorporate extensions to the statistical model. For example this approach can accommodate differences in

the expected within-site variability between historical sites and allow the assessment of a new site to be influenced not only by the number of samples, but also by the degree of variability seen in that site's monitoring data. Similar generalisations of the Confidence of Class approach, although possible in principle, would not be practicable with the existing spreadsheet-based methodology.

The extensions to the current DARES model explored by the Bayesian analysis (Models 2 and 3) have shown that the Probability of Class can change markedly according to the statistical assumptions. For example, the probability of 'Moderate or worse' typically falls by 10% or more in moving from Model 1 to Model 3 – and in all but one case (site 106) the choice of model determines whether or not the site fails the critical 95% PoM trigger. This demonstrates that the need to evaluate alternative statistical models (en route to adopting the most appropriate one) is not merely an academic nicety.

Conclusions / Recommendations

The exercise has illustrated the general principle that differences in the *underlying statistical model* and its associated assumptions are likely to have a much greater influence on data analysis results than whether the *statistical methodology* employed to fit the model is frequentist or Bayesian. One consequence of practical use is that there is no immediate and pressing need to adopt one approach to the exclusion of the other: either may be used depending on both the modelling circumstances and the statistical preferences of the tool developer.

In the longer term, however, this argument is less convincing, given the variety of complex statistical issues being discussed in relation to various WFD tools. These include: the development of monitoring programmes to obtain more reliable estimates of spatial and temporal components of variance; the growing need for analysis methods that can supplement future monitoring data with information from (relevant) historical data; and the long-running debate over how best to deal with spatial variability in extending site-based results to water body-wide assessments. Given the superior flexibility and greater intuitive appeal of the Bayesian approach, we believe there is a good case for a more comprehensive assessment of Bayesian methods and software, building on the foundations laid by the present exercise.

Acknowledgements

We would like to thank the DARES team and especially Dr Martyn Kelly (Bowburn Consultancy) for providing us with the historic river diatom data set for this work.

Contents

1 Introduction

1.1 Background

The Water Framework Directive (WFD) states that all European surface waters should achieve good ecological status by 2015. Ecological status is an expression of the quality of the structure and functioning of biological elements associated with surface waters, classified in Annex V. Ecological status should be assessed using a reference condition approach and classification tools based on five biological elements (EU 2000).

To help achieve comparability of monitoring systems between the different Member States, each Member State is required to express the results of their monitoring systems as ecological quality ratios (EQRs) for the purposes of classification of ecological status. It is also stipulated that the precision and confidence achieved by the monitoring results must be quantified.

In developing the techniques required to implement this system, the Environment Agency and SNIFFER have collaborated on related R&D projects that are investigating sources of uncertainty in the application of the classification tools and the implications for the reliability of the classification schemes. To date, this work on quantifying the uncertainty in WFD water body assessment has mostly focused on a 'classical' or 'frequentist[1]' statistical approach. This has led to the Confidence of Class (CofC) method as described by Ellis & Adriaenssens (2006) and subsequently used by many of the WFD tool developers.

In a follow-on project entitled: 'Uncertainty estimation for monitoring for each of the WFD biological classification tools – Further work on classification, uncertainty and variability aspects', the question was raised as to whether or not an alternative to the Confidence of Class approach could be the application of a Bayesian statistical approach. Accordingly, it was decided to select a WFD tool for which there was a good set of historical data and to use this dataset to demonstrate the Bayesian approach. The data would also be analysed using the existing CofC approach, thus enabling the two sets of results to be contrasted.

[1] The term 'frequentist' relates to a main-stream body of statistical thought developed in the early decades of the 20th century (although the word itself was only coined in mid-century). The central idea is that the probability of an event is the limiting value of the relative frequency of occurrence of that event over a number of repetitions of some well-defined random experiment. It may be that all possible outcomes can be enumerated. For example, if we roll two dice, there are just 6×6 = 36 possible outcomes. Suppose we are interested in the event 'total = 10'. This is achieved by three of the outcomes (4+6, 5+5 and 6+4) - and so the probability of the event is 3/36, or 1/12. Alternatively, there may be an indefinitely large number of possible experiments - such as the taking of a random sample at a particular point in a water body within some specified time and date window - and the event of interest may be 'NH4 concentration exceeds 1 mg/l'. The probability of this event is then defined as being the limiting value of the proportion of samples with NH4 exceeding 1 mg/l, as the number of samples gets very large.

1.2 The DARES diatom data

After some discussion between the Environment Agency and several of the tool developers, the river diatom tool DARES was selected as an appropriate test case. A comprehensive account of the DARES tool can be found in Kelly et al. (2007).

The historical dataset kindly made available by Martyn Kelly (Bowburn Consultancy) consisted of Trophic Diatom Index (TDI) data for 105 sites spanning 1992 to 2003. Between 6 and 20 samples were taken per site, with an average frequency of about 9. In addition to the actual TDI data, the expected TDI value for each site was also provided. Individual EQR values could be calculated by:

EQR = (100 - Actual TDI)/(100 - Expected TDI).

1.3 Scope of project

The primary objective of the project was to illustrate and contrast the principles underlying the frequentist and Bayesian approaches. Consequently it was agreed that we would restrict the analyses to a basic set of statistical models. In particular:

- we have not explicitly taken account of temporal variation (such as seasonality and longer-term trend). Temporal variability at a site is treated as a random component indistinguishable from short-term environmental variation and measurement error;
- we have ignored the important issue of spatial variability (especially relevant when more than one site in the water body is being sampled). We are regarding the *site* as the primary unit to be assessed; and
- we have assumed that the EQR class boundaries are *given quantities* with *no associated uncertainty*.

These and other complicating issues can of course readily be handled by statistical methods (whether frequentist or Bayesian), but their treatment would not especially illuminate the main objective and would in any case require more effort than was available for the present brief exercise.

1.4 Structure of report

Following this introduction, the report contains five main sections. First, Section 2 provides some essential background to the two statistical approaches we have applied. Next, Section 3 outlines the application of the existing frequentist approach to the DARES data[2]. Section 4 describes the corresponding Bayesian treatment of the data. We have duplicated the structure of these two sections, as closely as possible, to emphasise aspects that are common to both approaches and where the essential differences occur. However, as the report is principally concerned with the Bayesian approach to water body assessment, the Bayesian sections of the report include

[2] The material in Section 3 will be familiar to some readers, but we include it for the purpose of comparison.

comparisons and contrasts, where relevant, with the equivalent frequentist approach. (We felt that this was simpler than a stand-alone section later in the report that compared the two approaches.)

Section 5 broadens the discussion and presents a table summarising the pros and cons of the two approaches. Finally, Section 6 presents the conclusions and recommendations from the project.

2 Statistical background

2.1 Introduction

Suppose a water body (WB) is to be characterised by mean EQR, with the assessment applying to the WB as a whole over a three-year period. If we could sample the WB at a large number of locations on each of a large number of occasions over the three years, the resulting mean EQR would be very close to the 'true' population value.

In practice, of course, that is never possible. Any quantity or 'parameter' estimated from a set of samples will never be equal to the true value of that parameter, except by a lucky chance; and the discrepancy between them is generally known as 'sampling error'[3]. One of the main benefits of a statistical approach is the ability to *quantify the uncertainty* in the estimate for a parameter. This is fundamental to both the frequentist and the Bayesian approach. The mechanisms used to achieve this - the confidence interval in the frequentist case and the probability interval in the Bayesian case - are superficially similar. However, the assumptions behind their derivation and the interpretation of the intervals themselves, are fundamentally different. It is useful, therefore, in any comparison of the two approaches to start with some introductory discussion outlining their underlying principles - and that is the purpose of the following two sections.

2.2 Frequentist approach

2.2.1 The confidence interval

Suppose we take n samples at random from some specified population and use the mean of these values to estimate the population mean. The basis of the frequentist approach to quantifying sampling error is to visualise what would happen if we repeated this sampling exercise many times, each time taking n random samples and calculating the sample mean. Let us suppose we know that the values in the population are Normally distributed with standard deviation σ. (The Normality assumption is not essential to the general argument, but it keeps the details simple). From statistical theory we can say that, in the long run, 90% of the sample means will fall within $1.65\sigma/\sqrt{n}$ of the true mean.

The next step is to turn this statement round to say that, in the long run, the *true* mean will fall within $1.65\sigma/\sqrt{n}$ of the *sample* mean on 90% of occasions. (See the discussion and illustration in Annex B.)

[3] The term 'sampling error' is something of a misnomer, as it carries connotations of a mistake of some sort. But if a coin comes down tails 6 times in 10 spins rather than the expected 5, clearly this is not an 'error': it is simply an example of the random variability to be expected with any type of sampling.

In reality we never have the luxury of being able to repeat our sampling exercise an indefinite number of times: we have just one sample mean based on one set of n samples; but we nevertheless wish to make some statement about its uncertainty. To resolve this impasse, there now follows a sleight of hand (which some people call a 'confidence trick'). Starting with the statement:

"The true mean will fall within $1.65\sigma/\sqrt{n}$ of the sample mean on 95% of occasions"

we reword this to say that:

"We are 95% *confident* that the true mean falls within $1.96\sigma/\sqrt{n}$ of the observed sample mean."

This defines what is known as a *confidence interval*. Note that it is not a statement of probability. The true mean is a fixed (unknown) quantity. It either does, or does not, lie within $1.96\sigma/\sqrt{n}$ of the sample mean; there is no question of there being a '90% chance' of its doing so. Rather, we have to accept that this is the *definition of confidence*. It is a long-run probability of the statement being true - but is applied in the uneasy circumstances of our never having more that the one actual sample mean in our possession.

This may seem a somewhat bleak criticism of the confidence interval ('CI') approach. To counter it, therefore, here is a practically useful interpretation of the phrase 'in the long run'. Suppose we carry out a monitoring programme covering 100 WBs and for each WB we calculate the sample mean EQR and its 90% confidence interval. What we can then say is that about 90 of these 100 intervals will be correct - that is, they will bracket the true mean EQR for that WB.

2.2.2 Confidence of Class

In the context of WFD classification, a key concern is to be able to say how confident we are, on the basis of data from a monitoring programme, that the WB falls within any particular class. This measure, termed 'Confidence of Class' (CofC) calls for a somewhat unconventional application of the confidence interval concept that involves an inverted form of the usual calculation. The following two examples introduce the approach and illustrate how CofC is calculated.

Example 1

Suppose a WB is to be classified by taking the mean EQR, derived from some specified monitoring programme and comparing this with a set of predetermined class boundaries. Let us say that the sample mean falls within the Good range (0.60 to 0.80). Clearly our best estimate of the true class is 'Good', However, because of sampling error, there is the possibility that the site may truly be High, or Moderate (or even worse) and so we need to calculate the confidence that the site is truly Good.

Suppose the sample mean EQR is 0.70 and the 90% confidence interval by coincidence is [0.60 - 0.80]. This means we are 90% confident that the true mean EQR for the site lies in the range 0.60 to 0.80. But this is exactly the range defining Good! We can therefore be 90% confident that the class is Good. Moreover, because of the way confidence intervals are calculated, the confidence that the true mean falls *outside* the interval can be split equally between the low and high ends. We can therefore be 5% confident that the site is truly High and 5% confident that it is Moderate or worse.

Example 2

A more realistic example is described in Annex A, showing how the 'repeated sampling' paradigm can be visualised in the context of CofC. It is useful to summarise the outcome here, as it illustrates how CofC calculations are done in practice.

The situation is similar to that described in Example 1, except that now we imagine obtaining a sample mean EQR of 0.717, with standard error 0.095[4]. This means that a 90% confidence interval would be 0.717 ± 1.65×0.095, namely [0.56 - 0.87]. As this is wider than [0.60 - 0.80], we can say at once that CofC(Good) is less than 90%. But by how much?

To quantify this, we have to perform two calculations. First, we 'tune' the confidence level so as to make the *lower* confidence limit *exactly equal to 0.60*. This leads us to a 78% CI of [0.60 - 0.83]. From this, we can say that the confidence of being below 0.60 is (100 - 78)/2 = **11%**.

Secondly, we determine the confidence level that makes the *upper* confidence limit *exactly equal to 0.80.* This gives us a 62% CI of [0.63 - 0.80], from which we can say that the confidence of being above 0.80 is (100 - 62)/2 = **19%**.

Finally, it follows that the confidence of being *inside* [0.60 - 0.80] must be 100 - 11 - 19 = **70%**.

2.3 Bayesian approach

2.3.1 Introduction

In recent years, Bayesian statistics has become a common approach to data analysis in many areas of science. However, Bayesian statistics is not new – it was first described in 1763 by the Rev. Thomas Bayes, an English vicar. The reason for this recent rise in popularity is due to the availability of powerful computers which can readily cope with the complexity of the Bayesian calculations.

Most data analyses require the construction of a statistical model that describes the relationship between the observed data (e.g. individual TDI values) and the unknown model parameters (e.g. mean EQR). With frequentist statistics, the starting point for data analysis is the data set itself, from which the unknown parameters are estimated. However, with Bayesian statistics, the starting point for data analysis is the prior belief about the unknown parameters, which are expressed in the form of 'prior probability distributions'. The data is then used to modify the prior distributions for each unknown parameter, to give 'posterior probability distributions'. This use of data to transform prior distributions into posterior distributions is governed by Bayes' Theorem:

$$\boxed{\text{Prob (parameter given data)} \propto \text{Prob (data given parameter)} \times \text{Prob (parameter)}}$$

or, in more formal statistical language,

[4] We assume that the standard error has been estimated from historical data with a large number of degrees of freedom, so that there is no need to introduce the t distribution in calculating the CI.

Posterior distribution of parameter $\propto$ Likelihood of data x Prior distribution of parameter

Bayesian analysis can therefore be thought of as a process of using data to update our belief in the values of model parameters. This process of updating (achieved by a chain of Bayesian analyses that sequentially improve our assessments) can be used utilising historical data to improve our assessment of new data (see the next section).

2.3.2 The prior distribution

The first step in the Bayesian analysis is to choose prior probability distributions for the unknown model parameters. In many practical situations, there is no prior information on the parameters of interest and in this situation so-called ‘uninformative’ prior probability distributions are used. These prior distributions are a compulsory component of the Bayesian approach, but they will have minimal influence on the data analysis. However, it will often be desirable to use ‘informative’ prior distributions, which ideally should be objectively derived from a data analysis. Where this is not possible, informative prior distributions could be derived from expert opinion or from the scientific literature.

Consider a site with a single observed TDI[5] of 60 and an expected TDI of 36.6. The ‘face value’ estimate of the EQR for this site is:

EQR = (100 – Actual TDI)/(100 – Expected TDI) = 40/63.7 = 0.628.

A Bayesian estimate of the mean EQR for a site from this one sample requires specification of prior distributions for the mean EQR of the site and the temporal variability of samples around this mean. For the purpose of describing prior distributions, we will only consider the mean EQR; further details of the variance will be described in Section 4. A possible prior distribution for the mean EQR could be derived from an analysis of historic data, to obtain the overall frequency distribution of the mean EQRs at sites throughout rivers in England and Wales. This prior distribution for the mean EQRs has a mean of 0.659 and is shown by the thin continuous curve in Figure 1.

[5] The Bayesian analysis of the DARES data described in Section 4 keeps the actual and expected TDI scores as distinct variables, rather than modelling the EQR directly.

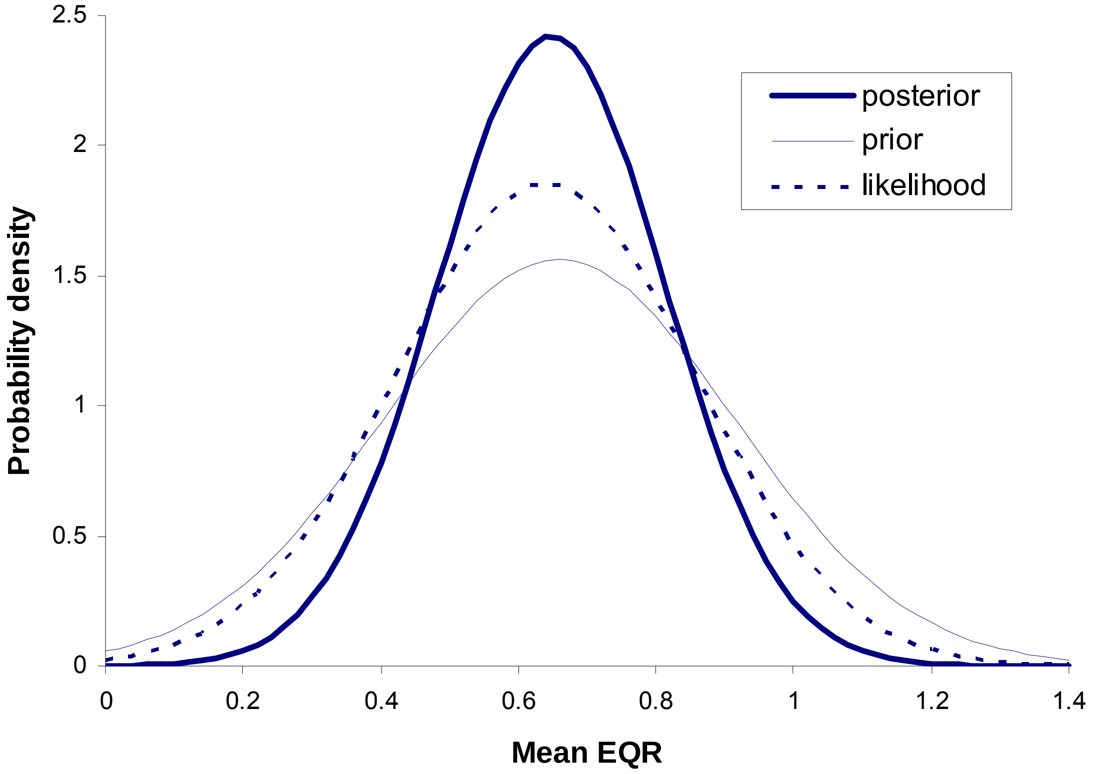


Figure 1 **Prior distribution, likelihood of data and posterior distribution for the mean EQR at a site with a single observed TDI of 60 and an expected TDI of 36.3.**

2.3.3 The likelihood function

The likelihood function describes the information about the model parameters contained in the data set. Whilst the term may not be often encountered in applied data analysis, it is at the heart of most frequentist methods, which are based on the idea of “maximum likelihood”. For example, the likelihood function peaks at 0.628 (the dotted curve in Figure 1) and so the maximum likelihood estimate of the mean EQR is 0.628 – the same as the face-value estimate calculated above. The concept of ‘likelihood’, as distinct from ‘probability’ and ‘confidence’, will not be described further in this report.

From Bayes’ Theorem, the posterior distribution is proportional to the product of the prior distribution and the likelihood function (Figure 1). The posterior distribution can therefore be thought of as the combination of the information about a parameter contained in the data (likelihood function) and the prior knowledge about the parameter (prior distribution).

In many situations, the data will provide considerable information about the unknown parameter and will be the dominant influence on the posterior distribution. Conversely, in situations where the data values are few or noisy and there is strong prior information, the prior distribution will dominate the posterior distribution. In this example, with just a single sample, neither the prior nor the likelihood provide much information about the mean EQR and the posterior distribution covers a wide range of possible values. However, the variability of the posterior distribution is less than that of the likelihood function. Thus a Bayesian analysis that is based on informative prior information will be more precise than the corresponding frequentist analysis (based on the likelihood function alone).

2.3.4 The posterior distribution

The posterior distribution provides all of the information available about a parameter from the Bayesian analysis, from which other summary statistics can be derived, such as the mean, median, mode, standard error, or probability of class (PofC). In this example, the mean of the posterior probability distribution is 0.648, which is slightly higher than the face-value estimate of 0.628, due to the influence of the informative prior distribution.

With frequentist statistics, all parameters, such as the mean EQR of a site, are regarded as fixed, unknown constants. However, the Bayesian posterior distribution gives us the probability that the parameter will take a particular value, conditional on the data collected. It is possible with Bayesian statistics to talk about the probability that a mean EQR is above some threshold or between two class boundaries. It is not possible – or indeed meaningful – to do this with frequentist statistics, where parameters are viewed as constants.

A Bayesian 95% probability interval for a parameter is defined as an interval that encloses 95% of the posterior probability distribution for that parameter (see Figure 2). So we can state that there is a 95% probability that the mean EQR lies within the interval. With frequentist statistics, in contrast, we state that on repeated sampling the confidence interval will enclose the true parameter value for 95% of data sets (See Annex B). The basis for the Bayesian probability interval is the inverse of the frequentist confidence interval: with Bayesian intervals, the parameter is viewed as a variable and the data set is constant, whereas with frequentist confidence intervals, the parameter is viewed as a constant and the data set is a variable. This inversion of the (frequentist) probability of data given a parameter, to the (Bayesian) probability of a parameter given data, is what has been achieved by the application of Bayes' Theorem.

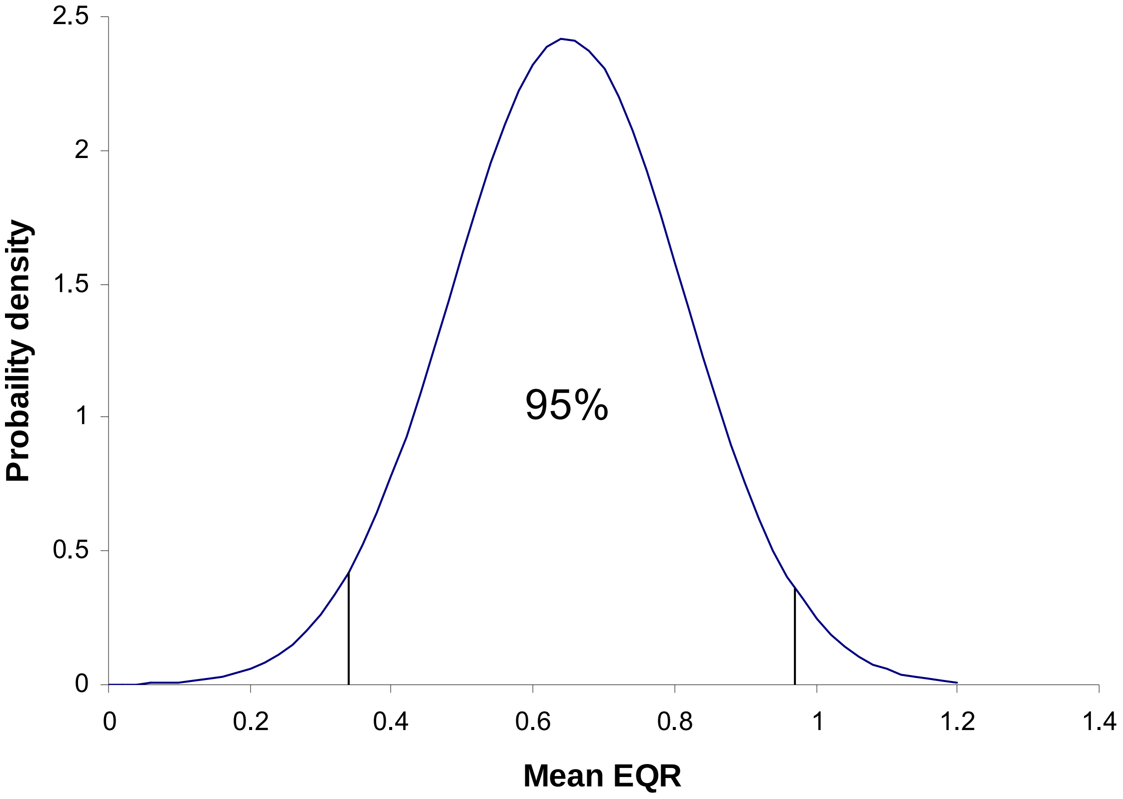


Figure 2 Bayesian posterior distribution, showing 95% probability interval, for the mean EQR at a site with a single observed TDI of 60 and an expected TDI of 36.3.

The Bayesian approach leads directly to the calculation of PofC. For any specified class this is defined as the probability that the unknown parameter falls within the class and is calculated from the proportion of the posterior probability distribution that falls between the two relevant class boundaries. The boundaries for the high, good, moderate, poor and bad EQR classes are 0.93, 0.78, 0.52 and 0.26 respectively and the probabilities of class are shown in Figure 3.

This illustrates one of the advantages of Bayesian statistics, by using data to modify a (prior) probability distribution, the output from Bayesian data analysis is also in the form of a probability distribution. This provides a very simple and intuitive way of summarising the uncertainty associated with the data analysis – something that is not so readily done with frequentist analysis. Compare the intuitive description of Bayesian PofC with the more complex description of frequentist CofC in Section 2.2 and Annex A. Indeed, many descriptions of CofC incorrectly describe the approach in terms of *probability* of class, thereby inadvertently switching to the more natural Bayesian definition!

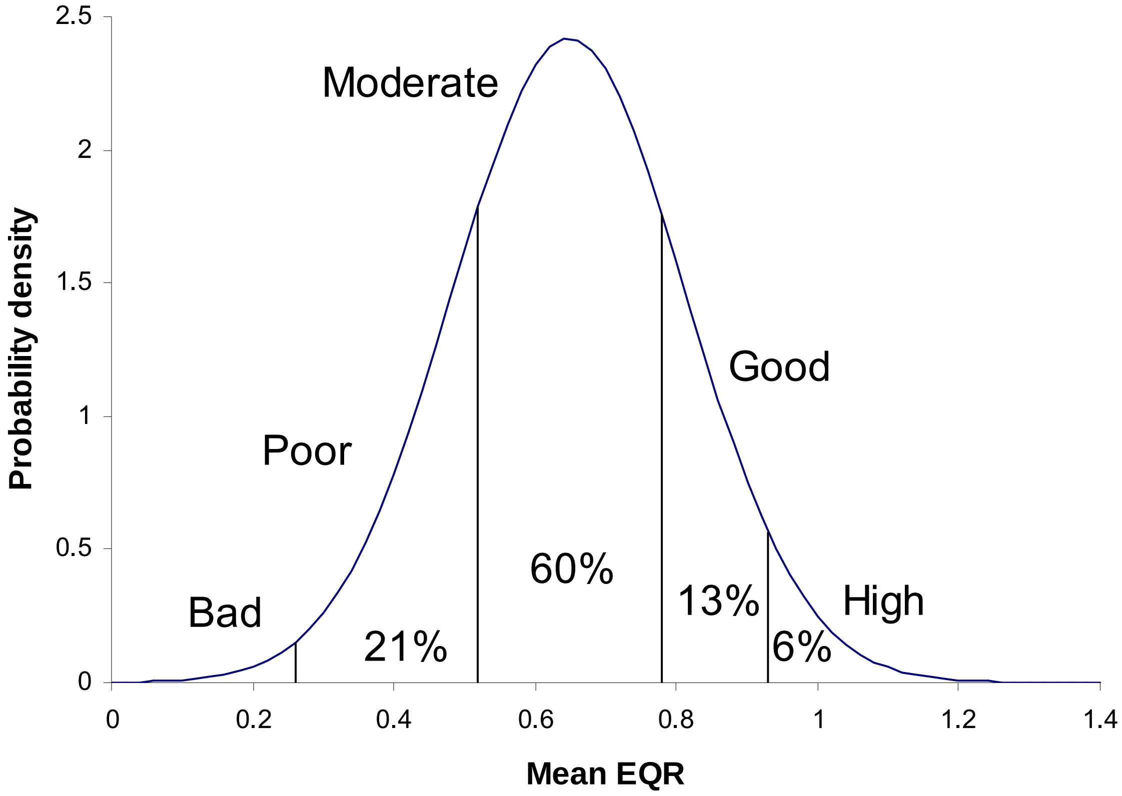


Figure 3 **Probability of class for the mean EQR at a site with a single observed TDI of 60 and an expected TDI of 36.3.**

3 Applying the frequentist approach

3.1 Statistical model

Notation

$TDI_{i,,j}$ = Observed value of TDI for the j-th sample at site i.

$ExpT_i$ = Expected value of TDI at site i (calculated from reference sites model)

$EQR_{i,j}$ = Value of EQR for the j-th sample at site i

= $(100 - TDI_{i,j})/(100 - ExpT_i)$.

Model

Given our initial decision to work with the aggregated DARES data, we had two main modelling options. One was to model TDI and ExpT separately and hence derive a model for $(100 - TDI_{i,j})/(100 - ExpT_i)$. The other option was to model the EQR ratio directly. Given that we had little information about the error structure of ExpT – and that its standard error was in any case likely to be small in relation to the variability shown by TDI – we decided to adopt the latter option.

The simplest model we can fit to the DARES data is as follows:

$EQR_{i,j}$ = $\mu + \alpha_i + e_{i,j}$, where:

μ = the overall mean;

α_i = effect (i.e. deviation from the overall mean) for site i; and

$e_{i,j}$ = deviation from model at sampling occasion j due to temporal trends at the site, random environmental variability and measurement error.

We also assume that the $e_{i,j}$ values are Normally distributed, with mean zero and standard deviation σ_i.

Note that, to keep the analysis as simple as possible, we have allowed the within-site temporal variation to be absorbed within the $e_{i,j}$ term. This is not too crude a simplification: unless there is a marked year-to-year trend applying across most or all of the 105 sites, a substantial part of the temporal variability would actually be spatial-temporal interaction, in which event this would be confounded with the $e_{i,j}$ term anyway.

We further assume that σ_j , the site-specific standard deviation term, is related to the site mean μ_j by the following relationship:

σ_i = $A + PX + QX^k$, where X is the mean for site i (i.e. $X = \mu + \alpha_i$).

We also impose the further constraints that the end points of the curve - that is, where X = 0 and X = 1 - are 'anchor points' specified by the user, having regard to the measurement error to be expected in these limiting cases. This is equivalent to specifying A (the predicted value of σ at X = 0) and A + P + Q (the predicted value of σ at X = 1).

Future assessments

For any future WB assessment, we assume for simplicity that:

- the WB will be sampled at just one site;
- samples will be taken randomly through time;
- the number of samples (n) will be too small to allow a site-specific standard error to be calculated and so a default value of s/√n will instead be used (where s is the predicted standard deviation for the observed mean EQR at the site).

The associated CofC values can then be calculated as indicated in Section 2.2.2, making the assumption that the within-site variability is Normally distributed.

3.2 Software

Three Excel tools were circulated in mid-2006 at the end of the 'quantifying uncertainty' project:

CAVE (Combines Appropriate Variance Estimates)

This assumed that the user had previously obtained various relevant components of variance (spatial, temporal, random environmental and measurement error) and now wished to combine them to obtain an appropriate uncertainty measure for WB assessment. The purpose of CAVE was to illustrate how this task depended critically on (a) the proposed monitoring programme and (b) the monitoring objective.

SDvMean

This assumed that the user had replicated or time-series EQR data at each of a number of sites and that the variability at each site was representative of that which would apply in future WB assessments. SDvMean provided a method for fitting an 'upturned wok' curve to a plot of standard deviation at a site to mean EQR at the site - the idea being to reflect the tendency for the standard deviation to fall off towards zero for very poor and very good sites.

CofC

Finally, CofC took the model determined by SDvMean, plus information on the number of planned samples per site and the EQR class boundaries and produced two plots - one showing CofC curves for the five WFD classes and another showing the Risk of Misclassification as a function of the true mean EQR.

Limitations of the current software

It should be emphasized that the above Excel tools were made available to the WFD tool developers primarily as a working illustration of how the recommended approach could be applied in certain specific circumstances, rather than purporting to be definitive software and they were accepted in that spirit. It was not intended that the tools would be able to cope with extensions or generalisations of the statistical model that incorporated, for example, further spatial or temporal components of variability or allowed for a site-specific variance component. Developments of that sort would call for the use of proper statistical software rather than spreadsheet-based analysis.

3.3 DARES results

3.3.1 Standard deviation model

In the present example we have no need of CAVE, as our starting point is the set of EQR means and standard deviations for the 105 sites. We can therefore proceed immediately to the SDvMean spreadsheet. This produces the scatter diagram shown in Figure 4. (Note that six of the mean EQR values were slightly greater than 1.0. As the tool expects all EQR values to fall within the range [0-1] we had to truncate these to 1.0.)

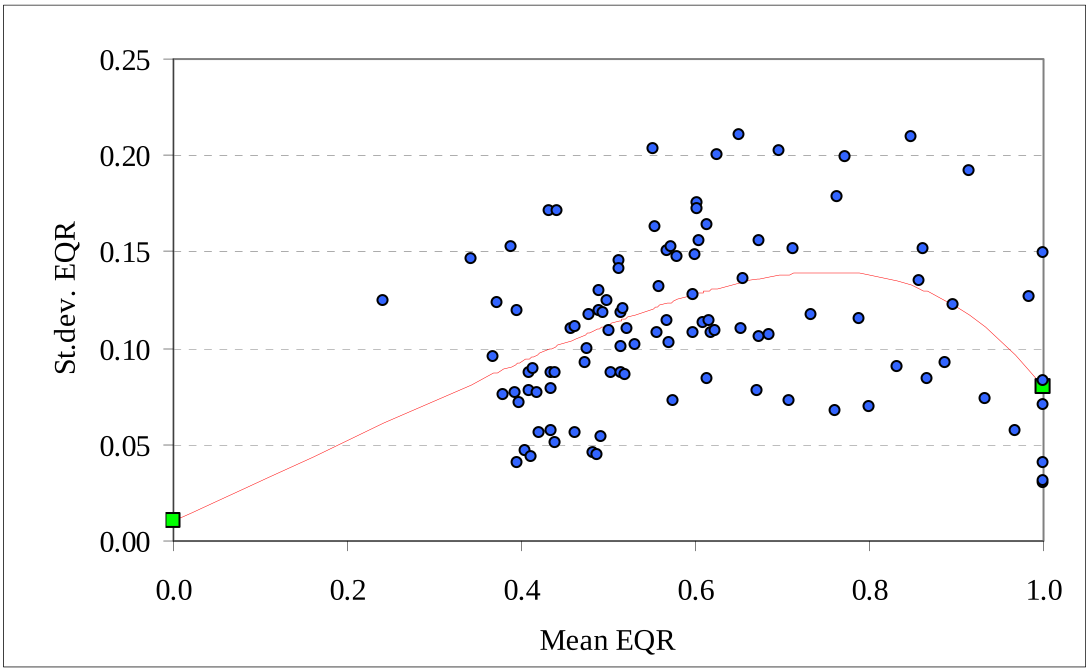


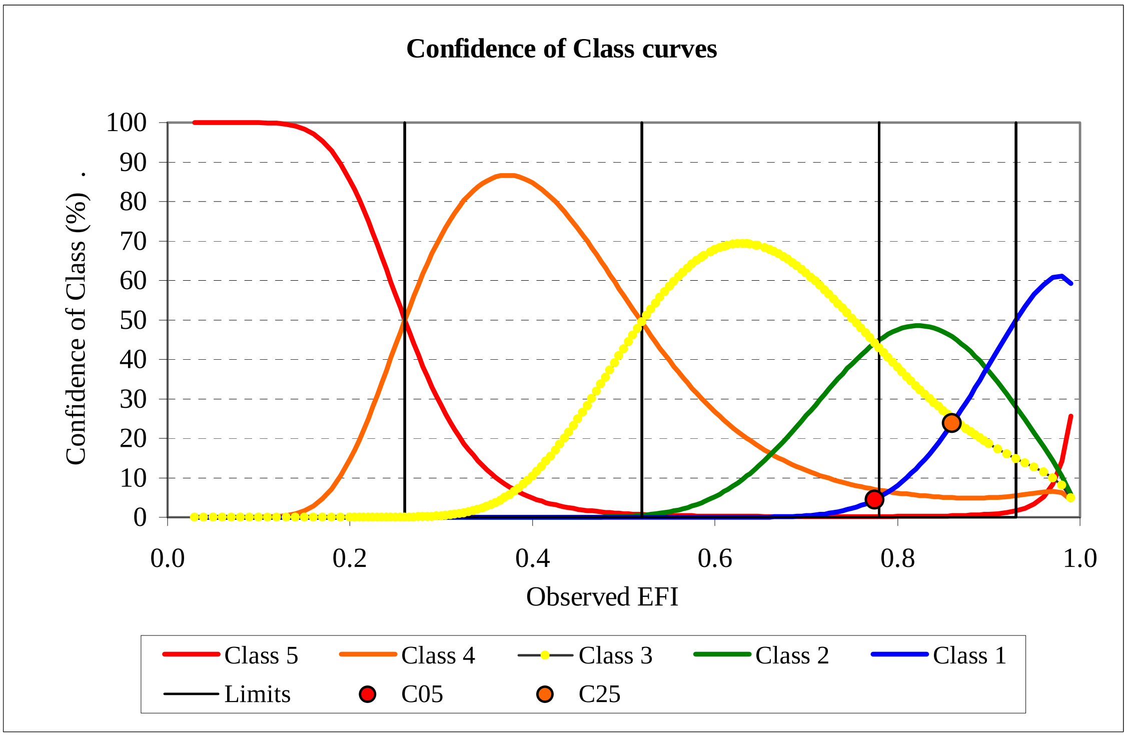
Figure 4 Plot of EQR standard deviation versus EQR mean for DARES sites

The plot also shows the standard deviation model that the tool fitted to the data, using anchor points of 0.01 and 0.08. We took the parameters of this model and plugged them into the CofC tool. We were then ready to calculate the CofC curves for any specified number of samples.

3.3.2 CofC results based on one sample at a site

Figure 5 shows the CofC curves that would apply if a DARES assessment were carried out on the basis of a single sample. The degree of discrimination is fairly good for sites at the poor end of the scale, but gets progressively worse as we move from left to right. For example, given an observed EQR of 0.4 we could be nearly 90% confident that the site was Poor, but given an EQR of 0.85 we would have less than 50% confidence that the site was Good and about 25% confidence both for Moderate and for High.

Note, the calculations go haywire at the extreme right-hand end of the plot because the artificially imposed limit of 1.0 is clearly inconsistent with the substantially non-zero standard deviations seen at the right-hand end of Figure 4. (This could be tidied up if necessary).



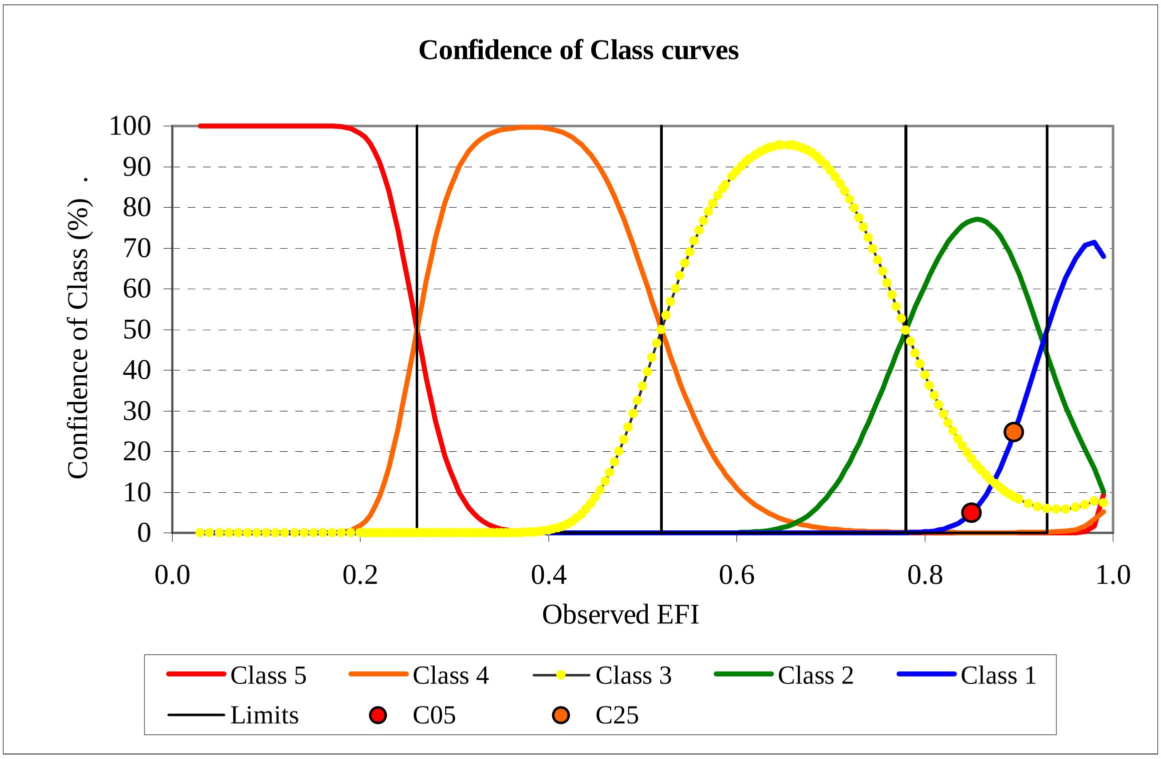
No of rep. samples:	1

C05	**0.775**
C25	**0.860**

Figure 5 CofC curves for DARES assessments based on one sample

3.3.3 CofC results based on four samples at a site

Figure 6 shows the CofC curves that would apply if a DARES assessment were carried out on the basis of four randomly selected samples over the monitoring period rather than just one. The standard error for a site is now only half of what it was before (i.e. s/√4) and as a consequence the degree of discrimination is much improved. For the examples cited earlier, the CofC values for the face-value class are now 100% for an EQR of 0.4 and 77% for an EQR of 0.85 (with the confidence of either neighbouring class down to about 11%).



No of rep. samples:	4

C05	**0.850**
C25	**0.895**

Figure 6 CofC curves for DARES assessments based on four samples

3.3.4 Application of CofC to new sites

Suppose we have seven additional DARES sites, each with the same expected TDI of 36.3 and with the observed TDI data shown in Table 1. The ‘face value’ estimate of the EQR for each of the sites is:

EQR = (100 - Mean TDI)/(100 - Expected TDI) = 40.0/63.7 = 0.628.

Thus the seven sites all fall a little below the centre of the Moderate class (for which the boundaries are 0.52 and 0.78).

Table 1 Illustrative TDI data for seven new sites

Site	Observed TDI values	Mean TDI
106	60	60
107	60, 60, 60	60
108	60, 60, 60, 60, 60, 60	60
109	50, 60, 70	60
110	50, 50, 60, 60, 70, 70	60
111	40, 60, 80	60
112	40, 40, 60, 60, 80, 80	60

For Site 106, the EQR is based on n = 1 samples, so we can determine the CofC values directly from Figure 5. The results are as shown in the first row of Table 2 below. In particular, we can be 69.3% confident that the site is Moderate.

Sites 107, 109 and 111 are all based on n = 3 samples, so their CofC values can be obtained from Figure 6. For these sites, the confidence of Moderate has risen to 90.4%.

Finally, we can obtain the results for Sites 108, 110 and 112 from the equivalent CofC graph for n=6. We see that the confidence of Moderate has now risen to 97.3%.

Table 2 Confidence of Class for seven new sites

Site	Confidence of Class (%)				
	Bad	Poor	Moderate	Good	High
106	0.3	21.2	69.3	9.3	0.0
107	0.0	8.5	90.4	1.1	0.0
108	0.0	2.6	97.3	0.1	0.0
109	0.0	8.5	90.4	1.1	0.0
110	0.0	2.6	97.3	0.1	0.0
111	0.0	8.5	90.4	1.1	0.0
112	0.0	2.6	97.3	0.1	0.0

4 Applying the Bayesian approach

4.1 Introduction

In the analysis of the DARES data using the frequentist approach, the historical data from 105 sites is first analysed to quantify the level of temporal variability using the SDvMean spreadsheet (Section 3.3.1).CofC curves are then calculated for different numbers of samples (Sections 3.3.2 to 3.3.3). This information is applied to a new data set of seven sites for which classification errors need to be calculated (Section 3.3.4).

A similar two-stage analysis of the two data sets could be employed with the Bayesian approach (Figure 7a). In the first stage of the analysis, a Bayesian analysis of the historical data would produce the posterior distributions for the overall mean EQR and the within-site variability. As we have no prior expectation regarding these parameters, we would use uninformative prior distributions. Then in the second stage of the analysis, the posterior distributions from the historical analysis are used as *informative prior* distributions for the analysis and classification of the new data. This two-stage Bayesian analysis is, therefore, similar to the two-stage frequentist analysis, except that the Bayesian analysis transfers information on the variability of EQRs via a probability distribution (rather than a single value from Figure 4) and transfers information on both the variability and mean EQR values.

Whilst this two-stage Bayesian analysis would help to contrast the Bayesian and frequentist approaches, in practice the Bayesian approach would probably be undertaken in a single step with all of the data analysed in a single model (Figure 7b). This means that in practice, only uninformative priors need to be specified. However, the Probability of Class (PofC) values estimated from these two procedures will be nearly identical.

In contrast, the analysis of all available data, both past and present, in a single model is not possible with the frequentist approach described in Section 3 and it would be quite a complicated matter to extend the approach so that it could do so.

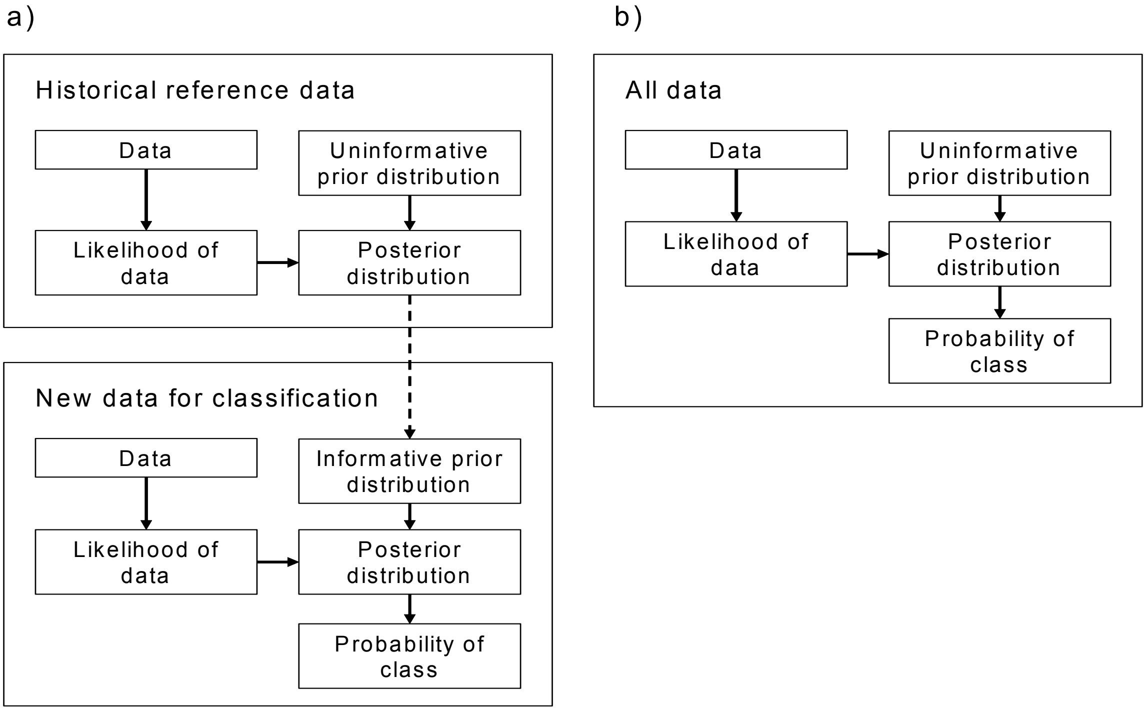


Figure 7 **Bayesian analysis of DARES data using a) a two-stage approach and b) a one-stage approach.**

4.2 Statistical model

As with the frequentist approach (Section 3.1) the Bayesian analysis is based on a statistical model. The most basic model for the Bayesian approach (Model 1) is:

Bayesian Model 1

$\text{logit}(\text{TDI}_{ij}/100) \sim \text{Normal}(\alpha_i, \tau)$

$\alpha_i \sim \text{Normal}(\mu_\alpha, \tau_\alpha)$

$\text{EQR}_i = (1 - \text{logit}^{-1}(\alpha_i)) / (1 - \text{ExpT}_i/100)$

where

TDI_{ij} is the TDI of the jth sample at the ith site,

α_i is the mean (logit transformed) TDI at the ith site,

τ is the temporal variability in TDIs within sites (measured as 1/variance)

$_\alpha$ is the mean (logit transformed) TDI across all sites,

τ_α is the spatial variability between sites (measured as 1/variance)

ExpT_i is an estimate of the expected TDI at the ith site under reference conditions and

EQR_i is the EQR at the ith site.

This is very similar to the model underlying the frequentist approach except that here, for reasons that we explain shortly, we have chosen to model the TDI (expressed as a proportion) rather than the EQR. (This makes practically no difference as far as the later comparisons with the frequentist results are concerned.) As TDI/100 must lie between 0 and 1, it is appropriate to model it on a logistic scale. Accordingly, we assume that logit(TDI_{ij}/100) varies randomly according to a Normal distribution with mean α_i and 1/variance[6] τ. This brings a useful benefit regarding the implied variance of the EQR. As logit(TDI/100) has a constant variance (1/τ) the TDI on an un-transformed scale will have a maximum variance at a TDI of 50 and zero variance at TDIs of 0 and 100. (See the discussion on this in Ellis and Adriaenssens (2006)). For data that has an expected TDI ($ExpT_i$) of around 36, therefore, the variance of the EQR will have a maximum at a mean EQR of 0.78 (= (100-50)/(100-36)) and a minimum at mean EQRs of 0 (= (100-100)/(100-36)) and 1.56 (= (100-0)/(100-36)). Thus, by modelling logit(TDI/100) the relationship between the variance of the EQR and the mean EQR naturally follows an 'upturned wok', similar to that in Figure 4, but without the explicit need either to set arbitrary anchor points,or to model the shape of the relationship.

A further difference between Bayesian model 1 and the frequentist model is that the former assumes not only that the logit(TDI_{ij}/100) varies randomly within a site according to a Normal distribution, with mean α_i and 1/variance τ, but also that the site means themselves vary randomly according to a Normal distribution with mean μ_α and 1/variance τ_α. Thus, there are two sources of random variation in this model, described by τ and τ_α.

One advantage of the Bayesian approach and the use of Bayesian software is that the models can easily be changed, if necessary, to better describe the data. For example, it was clear from the data that the variability within each site (1/τ) was not constant, with some sites exhibiting a greater degree of variation than others. We therefore decided that the degree of variability should itself vary from site to site. This led to Model 2, shown below, in which the changes from Model 1 are shown in **bold**:

Bayesian Model 2

$\mathrm{logit}(TDI_{ij}/100) \sim \mathrm{Normal}(\alpha_i, \boldsymbol{\tau_i})$

$\alpha_i \sim \mathrm{Normal}(\mu_\alpha, \tau_\alpha)$

$\boldsymbol{\log(\tau_i) \sim \mathrm{Normal}(\mu_\tau, \tau_\tau)}$

$EQR_i = (1 - \mathrm{logit}^{-1}(\alpha_i)) / (1 - ExpT_i/100)$

where

[6] For technical reasons WinBUGS expresses variability in terms of a quantity called 'precision', which is defined as 1/variance. Unfortunately this usage clashes with the frequentist meaning of 'precision', namely the half-width of a confidence interval - with the further complication that the two usages work in opposite senses. (A numerically large Bayesian precision is equivalent to a numerically small frequentist precision.) To avoid confusion, therefore, we will refer to precision in the WinBUGS sense as '1/variance'.

τ_i is the temporal variability in TDIs within the ith site (measured as 1/variance)

μ_τ is the mean (log transformed) variability across all sites and

τ_τ is the degree to which temporal variability varies between sites (measured as 1/variance).

In a final refinement (Model 3) the uncertainty associated with modelling reference conditions was included. To do this, we assumed that ExpT_i, the expected value of TDI at site i (which is calculated from the reference sites model) was an imprecise estimate of the truth (TrueT_i) and that the 1/variance associated with the reference sites model (τ_e) was assumed to be known. The model (with the refinements from the previous model again shown in **bold**) now becomes:

Bayesian Model 3

$\mathrm{logit}(\mathrm{TDI}_{ij}/100) \sim \mathrm{Normal}(\alpha_i, \tau_i)$

$\alpha_i \sim \mathrm{Normal}(\mu_\alpha, \tau_\alpha)$

$\log(\tau_i) \sim \mathrm{Normal}(\mu_\tau, \tau_\tau)$

$\mathbf{ExpT_i \sim Normal(TrueT_i, \boldsymbol{\tau_e})}$

$\mathrm{EQR}_i = (1 - \mathrm{logit}^{-1}(\alpha_i)) / (1 - \mathbf{TrueT_i}/100)$

where

TrueT_i is the true (unknown) expected TDI at the ith site under reference conditions and

τ_e is the error associated with the reference condition model (measured as 1/variance).

Model 3 has four unknown parameters (μ_α, τ_α, μ_τ, τ_τ) for which Bayesian prior distributions have to be specified. As previously mentioned, we chose to use uninformative priors in this analysis. This is appropriate when we have no prior knowledge of the parameters (e.g. from the literature) before analysing the data. For τ_α and τ_τ (which must be positive) uninformative Gamma distributions were used and for μ_α and μ_τ (which can be positive or negative) uninformative Normal distributions were used.

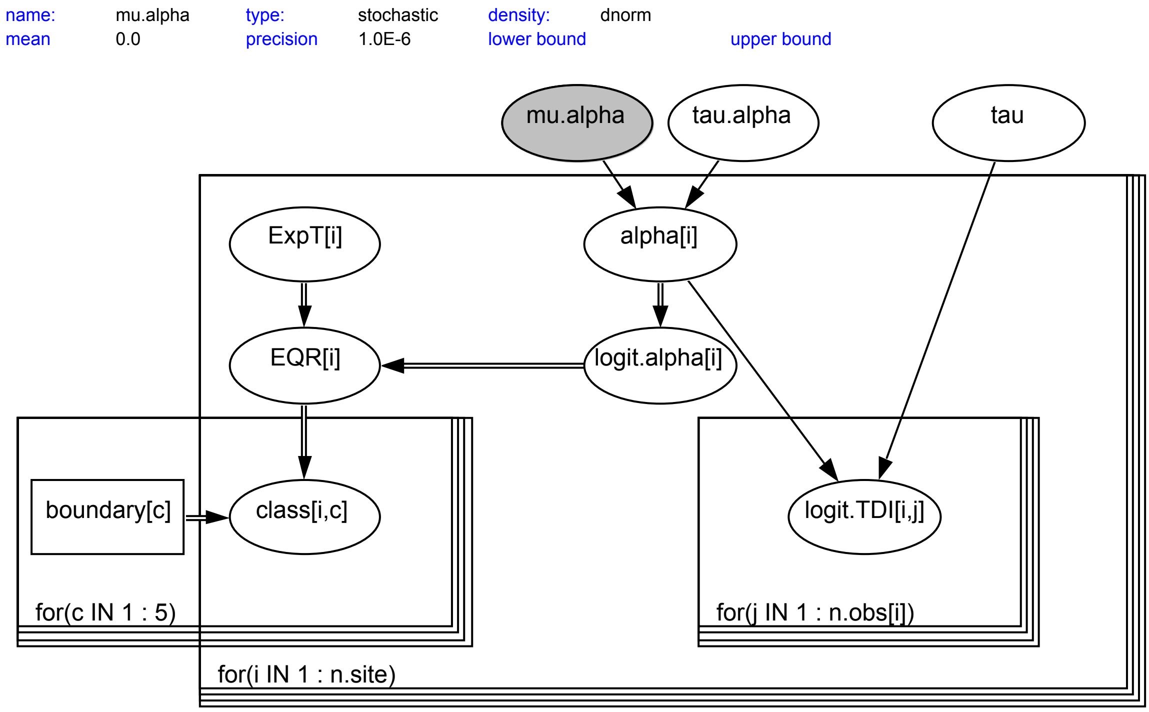
4.3 Software

Whilst the Bayesian approach is conceptually simpler than the frequentist approach, until recently it has been much more difficult to implement. However, with the increase of computing power, Bayesian statistics has become much more accessible, thanks to computer-intensive methods that use iterative calculations.

One such iterative method is 'Gibbs Sampling', which is used by the Bayesian statistical software package BUGS (**B**ayesian inference **U**sing **G**ibbs **S**ampling, Lunn et al. 2000). BUGS is freely available and the Windows version, WinBUGS 1.4.3, was used for this project (www.mrc-bsu.cam.ac.uk/bugs).

With WinBUGS, it is possible to specify models within the software just by drawing flow diagrams, known as 'doodles', with the mouse. The doodle for Model 1 is shown in Figure 8 (a screen-shot from the software). Each parameter of the model is represented by a node and the dependencies between the parameters are shown by arrows. Single arrows denote stochastic relationships – that is, relationships involving statistical uncertainty – whereas double arrows denote deterministic relationships. Repeated components of the model are enclosed within the large rectangles ('plates') with i denoting each site, j denoting each observation within a site and c denoting the five WFD classes.

The doodle corresponds exactly with the three equations given above for Model 1. Alpha (α_i) is stochastically linked to mu.alpha (μ_α) and tau.alpha (τ_α) TDI$_{ij}$ is stochastically linked to alpha (α_i) and tau (τ) and the EQR$_i$ is deterministically calculated from ExpT$_i$ and logit^{-1}(α_i). This doodle also shows the calculation of the PofC for each site (class[i,c]) given the class boundaries (held in the array 'boundary[c]'). The text at the top of the doodle gives details of the highlighted node – in this example, mu.alpha (μ_α). This is a parameter that requires an uninformative prior distribution and the details show that it is a stochastic node with a prior that is a Normal distribution (density is 'dnorm') with a mean of zero and a very low 1/variance (1.0E-6) i.e. a high variance. This distribution is therefore very nearly flat, which means that our prior belief is that all values, negative or positive, are equally likely.



name: mu.alpha type: stochastic density: dnorm

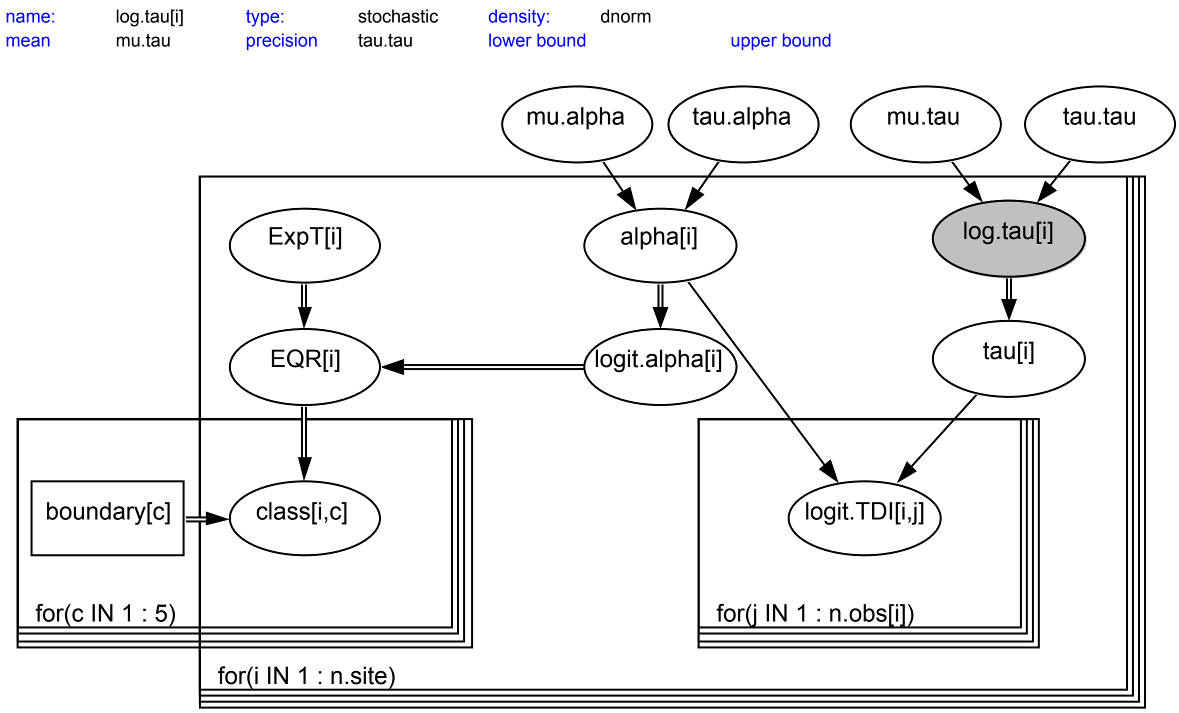
mean 0.0 precision 1.0E-6 lower bound upper bound

Figure 8 Doodle for Bayesian Model 1

To specify Model 2 in WinBUGS, the doodle is modified (see Figure 9) to show the between-site differences in variability. Log.tau[i] (that is, log(τ_i) in the algebraic model notation) is the highlighted node in this example, showing that it is now a stochastic node that follows a Normal distribution with mean mu.tau (μ_τ) and 1/variance tau.tau (τ_τ).

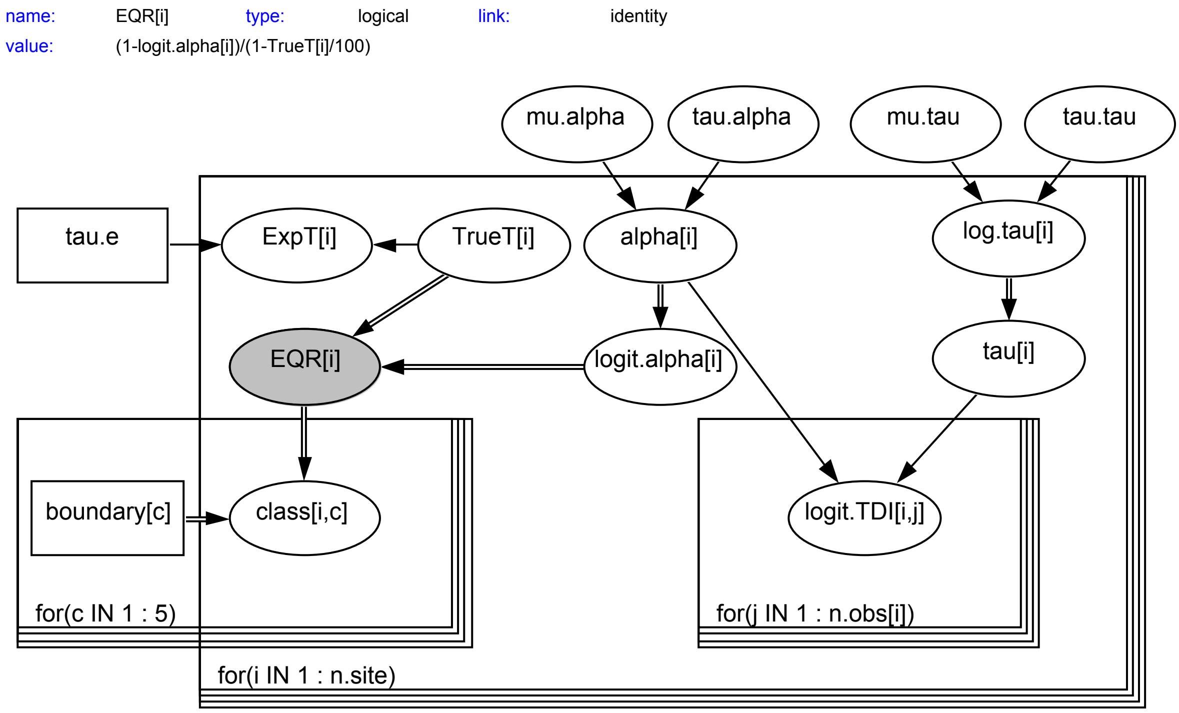
Model 3 (see Figure 10) adds one final refinement; it assumes that the expected TDI score (ExpT[i]) from the reference sites model is an imprecise estimate of the true expected TDI score (TrueT[i]) with 1/variance tau.e (τ_e). The highlighted node is EQR[i], showing that it is a deterministic node that is a function of the true, but unknown, expected TDI score (TrueT[i]).

Once the model has been specified as a doodle, the raw data is entered into WinBUGS. The complete data set comprises the observed TDI score for each site and occasion (TDI[i,j]); the expected TDI score for each site (ExpT[I]); and, for Model 3, the assumed level of error in the reference site model (tau.e).



name: log.tau[i] type: stochastic density: dnorm
mean mu.tau precision tau.tau lower bound upper bound

Figure 9 Doodle for Bayesian Model 2.


name: EQR[i] type: logical link: identity

value: (1-logit.alpha[i])/(1-TrueT[i]/100)

Figure 10 Doodle for Bayesian Model 3.

4.4 DARES results

4.4.1 Estimates of model parameters

Model 1 has three primary unknown parameters (μ_α, τ_α, τ) and Models 2 and 3 have four (μ_α, τ_α, μ_τ, τ_τ). The uncertainty associated with the reference site model (τ_e) was assumed to be 0.01. The Bayesian estimates for these parameters are given in Table 3.

Table 3 Results for the three Bayesian models

Parameter	WinBUGS node	Estimate	Std error	95% probability interval	
				Lower	Upper
Model 1					
μ_α	mu.alpha	0.401	0.0618	0.281	0.523
τ_α	tau.alpha	2.531	0.3589	1.883	3.284
τ	tau	6.716	0.3322	6.075	7.380
Model 2					
μ_α	mu.alpha	0.401	0.0608	0.282	0.519
τ_α	tau.alpha	2.554	0.3623	1.896	3.306
μ_τ	mu.tau	2.081	0.08169	1.925	2.247
τ_τ	tau.tau	2.738	0.7876	1.568	4.622
Model 3					
μ_α	mu.alpha	0.401	0.0614	0.280	0.523
τ_α	tau.alpha	2.551	0.3602	1.903	3.316
μ_τ	mu.tau	2.081	0.0813	1.924	2.241
τ_τ	tau.tau	2.733	0.7946	1.562	4.627
τ_e	tau.e	0.01*	0	N/A	N/A

* Assumed value.

These results show that the three models are very similar in terms of the primary parameters. The low values for the parameter τ_τ suggest that there is a high degree of between-site variability in the within-site variances, confirming that Model 2 is a more realistic description of the data than Model 1. The Bayesian approach provides estimates of the uncertainty associated not only with the means but also the variances, and these uncertainties are all reflected in the final estimates of PofC.

4.4.2 Application of model to historic sites

The mean EQR values for the 105 historic reference sites are shown in Figure 11. The differences between Models 1 and 2 are small, but the uncertainty associated with model 3 is considerably higher than for the other two models. It should be remembered that Model 3 currently uses an assumed value for the uncertainty associated with the reference model. Even so, this result illustrates the potential influence of this source of uncertainty.

4.4.3 Application of model to new sites

Using a Bayesian approach, the PofC for new sites will vary according to both the *number* of samples contributing to the mean TDI value and the *variability* of those samples. In contrast, the CofC calculated using the frequentist approach takes account of the first of these elements, but *not the second*. It has generally been assumed that there will typically be too few samples to allow us to estimate the variability solely from the new data and so we have fallen back on the ‘upturned wok’ plot from historical data, showing the typical variability in EQR as a function of mean EQR at the site.

With the Bayesian approach, however, both the current variability (i.e. the likelihood of the variance parameter) *and* the historical variability (used to derive the prior distribution for the variance parameter) are utilised in the PofC calculation. The two extremes are:

- We have a lot of historical data and only a small amount of new data. In that case, the variability used in calculating PofC will be practically unchanged from the typical historical value (as with the CofC approach).

- We have limited historical data but a substantial amount of new data. In that case, the variability used in calculating PofC will be close to that obtained from the new data and only slightly modified by the historical evidence.

Most situations fall between these two extremes - and the statistical machinery of Bayes’ Theorem determines precisely how much weighting to give to the two components.

The PofC results for the seven sites described in Section 3.3.4, all of which have a ‘face value’ estimate of the EQR of 0.628, are shown in Table 4.

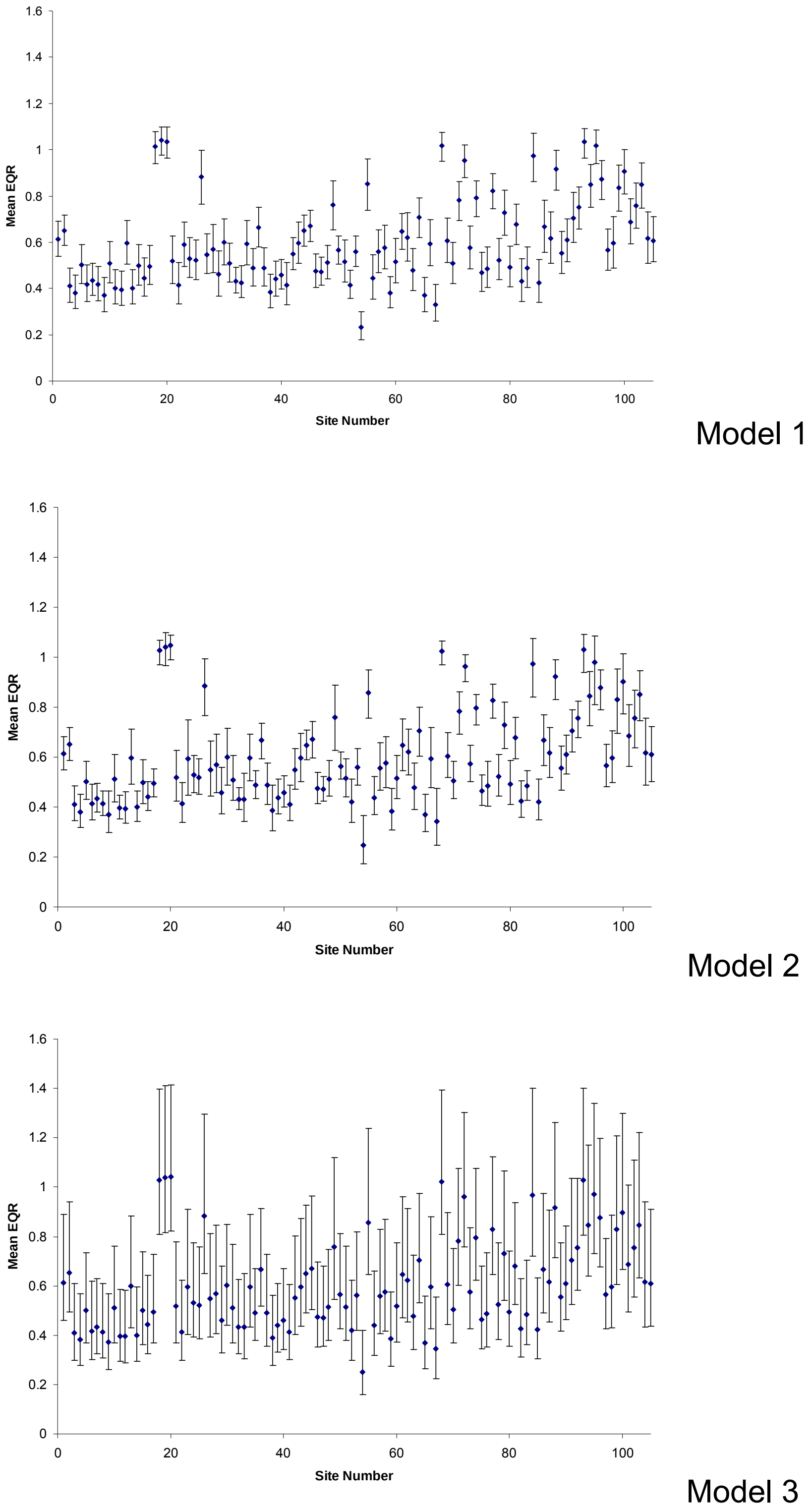


Figure 11 Mean EQR and 95% probability intervals for 105 sites

Table 4 Probability of Class for seven new sites

Site	Model	Probability of Class				
		Bad	Poor	Moderate	Good	High
106	1	0.000	0.184	0.698	0.110	0.009
	2	0.000	0.155	0.745	0.090	0.009
	3	0.001	0.213	0.597	0.134	0.055
107	1	0.000	0.076	0.892	0.032	0.000
	2	0.000	0.041	0.942	0.017	0.000
	3	0.000	0.146	0.716	0.107	0.031
108	1	0.000	0.026	0.970	0.004	0.000
	2	0.000	0.004	0.995	0.001	0.000
	3	0.000	0.113	0.770	0.097	0.021
109	1	0.000	0.086	0.887	0.027	0.000
	2	0.000	0.087	0.880	0.032	0.001
	3	0.000	0.173	0.687	0.109	0.031
110	1	0.000	0.032	0.964	0.004	0.000
	2	0.000	0.035	0.959	0.007	0.000
	3	0.000	0.145	0.734	0.098	0.023
111	1	0.000	0.120	0.860	0.020	0.000
	2	0.000	0.174	0.766	0.057	0.003
	3	0.000	0.234	0.617	0.109	0.039
112	1	0.000	0.056	0.942	0.002	0.000
	2	0.000	0.129	0.845	0.025	0.001
	3	0.000	0.214	0.655	0.102	0.029

Note: the shaded row corresponds to the example used in Section 2.3

From these results we see that, as expected, Site 106 has the greatest uncertainty (lowest probability of Moderate status) as the mean is based on only one sample. By comparing sites with the same variability but different sample sizes, we can see that sites with higher sample sizes (n=6) have lower uncertainty than sites with lower sample sizes (n=3). Furthermore, by comparing sites with the same sample size, but different variability, we can see that sites with the highest sample variability have the greatest uncertainty. This is in contrast to the frequentist approach (Table 2) where no account is taken of the variability in the sample data.

The Probabilities of Class also enable us to examine the differences between the three models in more detail. Whilst the PofC values from Models 1 and 2 are similar, Model 2 is more sensitive to the variability of the sample and less influenced by the variability of the 105 reference sites, when compared to Model 1. At sites with zero variability (107 and 108) Model 2 has the lowest uncertainty. At sites with high variability (111 and 112) Model 2 has the highest uncertainty; whereas at sites with intermediate variability (109 and 110) Models 1 and 2 are very similar. Model 3 builds on Model 2 by adding uncertainty in the expected TDI scores and, as a result, has greater uncertainty for all sites.

It is sometimes necessary to classify sites or water bodies where there are no data. In this situation, a Bayesian analysis enables classification to be based on the prior distribution, rather than the posterior distribution. For example, if a site has had no sampling then the likelihood of the data cannot be calculated and the prior distribution (Figure 1) provides probabilities of class (from Bad to High) of 0.030, 0.295, 0.381, 0.146 and 0.147 respectively. The probability of Moderate status is now only 0.381; as would be expected, the probabilities of class are much less precise when there are no data!

5 Discussion

5.1 Comparison of the two approaches

Bayesian Model 1 is the one most similar to the frequentist model and so it is instructive to compare the frequentist CofC values from Table 2 with the Model 1 PofC values from Table 4 (converted to percentages). This has been done in Figure 12.

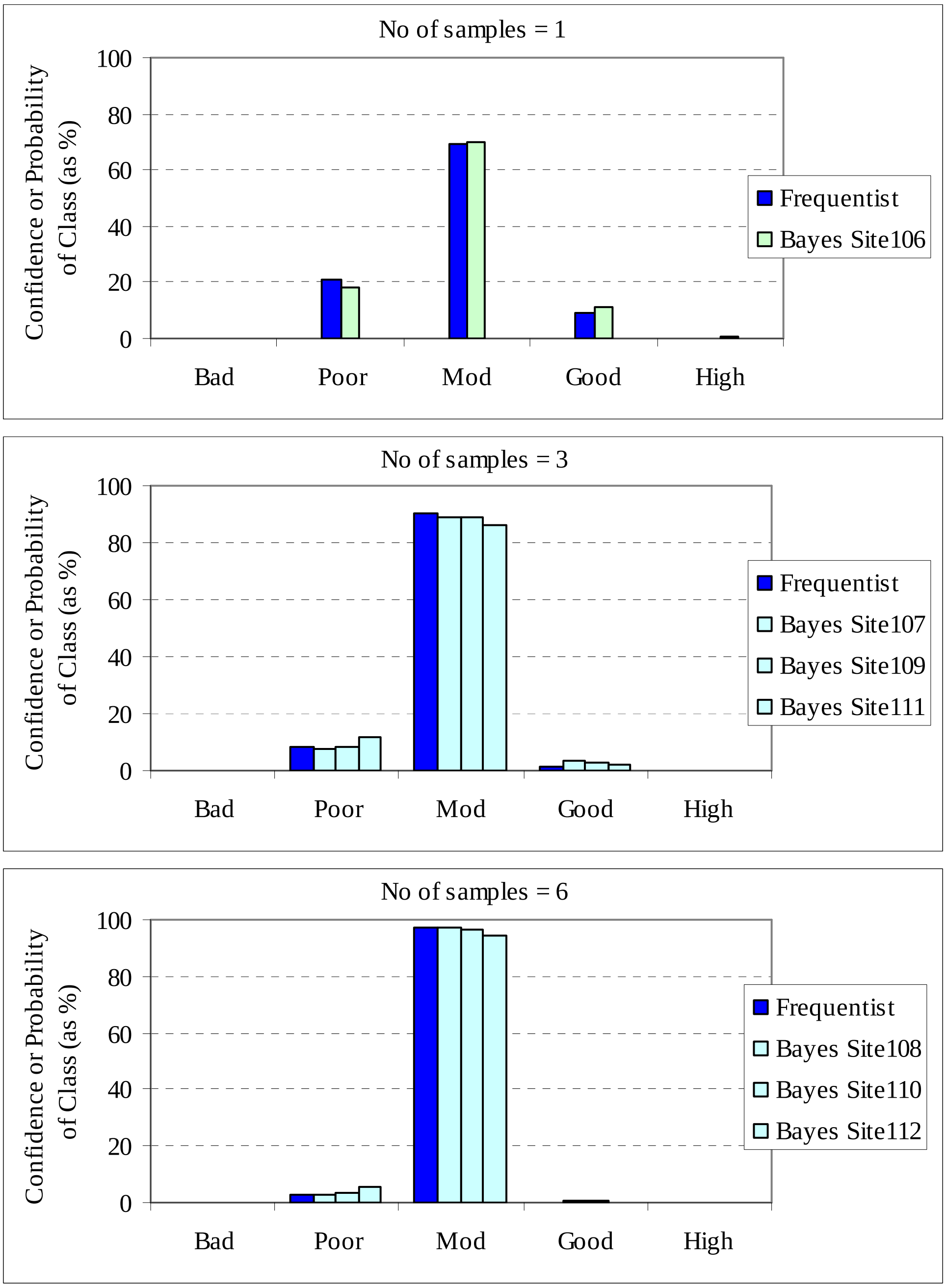


Figure 12 Comparison of frequentist and Bayesian results for seven new sites

It can be seen that the two sets of results are numerically in remarkably close agreement. This may appear to be something of a paradox, given the profoundly different philosophy underpinning the frequentist and Bayesian approaches. However, it is simply an illustration of the general principle (readily acknowledged by most statisticians with an unbiased foot in either camp) that, provided the same statistical model is fitted in each case, there will commonly be relatively little difference between the two sets of results unless there is good reason to use a very strong prior. In this example, a moderately strong prior from the historical data was used for classifying the new sites, but it appears that the increase in precision obtained has largely been offset by the more realistic assessment of uncertainty achieved by the Bayesian approach.

A more general comparison of the frequentist and Bayesian approaches, gathering together and summarising a number of points, identified. in earlier sections, is provided by Table 5.

5.2 Uncertainty in class boundaries

At the start of this report we stated that we would be assuming that the class boundaries were given quantities, with no associated statistical uncertainty. This is a convenient point to return to this issue and to discuss some of the consequences of relaxing this assumption.

Hamill & Ellis (2003) provided an illustrative Bayesian example to show how, given two overlapping reference sets of Good and High sites, these could be used to classify a new site without ever needing to set an (essentially arbitrary) Good/High boundary. It was perhaps unfortunate that this illustration immediately followed a description of the frequentist CofC approach in which the class boundaries were (as has traditionally been the case) assumed to be fixed quantities known without error. On a casual reading, therefore, It might have seemed that the marked difference in outcomes was attributable to some fundamental difference between the frequentist and Bayesian approaches, whereas in fact it was almost entirely due to the changed assumption about the class boundaries. The premises were different and so naturally the outcomes were different.

A particularly dramatic contrast was seen more recently when Wyatt (pers.comm, 2006) subjected the European Fish Index data to a Bayesian analysis. Wyatt used the five (heavily overlapping) sets of EFI reference sites as source data, rather than the derived class boundaries. This produced PofC curves that were much flatter and more widely spread across classes than the relatively sharply defined CofC curves previously reported for EFI by Adriaenssens and Ellis (2006). Again, it would have been incorrect to attribute these differences to the change in analytical method from frequentist to Bayesian. Rather, they were the consequence of a huge change in the assumptions made about the class boundaries – firstly that they were precisely known and secondly that they were very loosely defined because of the poor discrimination achieved by the sets of reference sites. A frequentist analysis that adopted the latter stance could be expected to produce similarly weak conclusions.

Table 5 Comparison of frequentist and Bayesian approaches for assessing WFD classification errors

Feature	Frequentist	Bayesian
Measure of classification errors	Confidence of Class (CofC)	Probability of class (PofC)
Interpretation of classification errors	Not straightforward (see Annex A) and easily misunderstood	Intuitive (see Figure 3)
Uncertainty in variance estimates	Ignored by present approach	Included in all models
Use of historic data	Used to estimate variances	Used to provide prior information for both means and variances
Variance – mean EQR relationship	Modelled empirically with 'up-turned wok' power curve	Included automatically within probabilistic structure of model
Variability of sample being classified	Ignored in CofC calculations	Included in PofC calculations
Uncertainty of classification	Determined by sample size alone (together with typical value of historical variability)	Bayesian approach based on informative priors from historical data, so more realistic.
Analysis of data from reference sites and new sites	Reference data set analysed first and results applied to new data for which CofC required	Reference data and new data can be incorporated into a single analysis
Software	A series of custom-designed spreadsheets	Bayesian software, e.g. WinBUGS (free)
Implementation	Requires a number of steps: 1) initial analysis, 2) SDvMean, 3) CofC	Analysis goes from raw data to PofC in a single step
Modification of models	Alternative models would require re-designing the spreadsheets, or graduating to proper statistical software	Alternative models can be simply explored by modifying the WinBUGS 'doodle'
Within-site variability	Assumed constant	Allowed to vary from site to site in Models 2 and 3
Error in reference site model	Ignored	Included in Model 3

6 Conclusions and Recommendations

A comparative analysis has been carried out for the DARES diatom tool using both a **frequentist** approach (as currently adopted by most of the WFD tools) and a **Bayesian** approach (as adopted by the Fisheries WFD tool).

We have used the two approaches to analyse illustrative data sets for each of seven 'new' sites. This has shown that, provided similar statistical assumptions are made in analysing the data for the 105 historical sites, there is very little practical difference between (a) the Confidence of Class values generated by the frequentist approach for the new sites and (b) the corresponding Probability of Class values arising from the Bayesian method.

More generally, the exercise has demonstrated the ease with which the Bayesian approach (allied to the WinBUGS software) can be adapted to incorporate extensions to the statistical model. For example, differences in the expected within-site variability between historical sites can be accommodated and the assessment of a new site can be influenced both by the number of samples and the degree of variability in that site's monitoring data. Similar generalisations of the Confidence of Class approach, although possible in principle, would not be practicable with the existing spreadsheet-based methodology.

The extensions to the current DARES model explored by the Bayesian analysis (Models 2 and 3) have shown that the Probability of Class can change markedly according to the statistical assumptions. For example, the probability of 'Moderate or worse' typically falls by 10% or more in moving from Model 1 to Model 3. In all but one case (site 106) the choice of model determines whether or not the site fails the critical 95% PoM trigger. This demonstrates that the need to evaluate alternative statistical models (en route to adopting the most appropriate one) is not merely an academic nicety.

The exercise has illustrated the general principle that differences in the *underlying statistical model* and its associated assumptions are likely to have a much greater influence on data analysis results than whether the *statistical methodology* employed to fit the model is frequentist or Bayesian. One practically useful consequence of this is that there is no immediate and pressing need to adopt one approach to the exclusion of the other: either may be used depending on both the modelling circumstances and the statistical preferences of the tool developer.

In the longer term this argument is less convincing, given the several complex statistical issues being discussed in relation to various WFD tools. These include: the development of monitoring programmes to obtain more reliable estimates of spatial and temporal components of variance; the growing need for analysis methods that can supplement future monitoring data with information from (relevant) historical data; and the long-running debate over how best to deal with spatial variability in extending site-based results to WB-wide assessments. Given the superior flexibility and greater intuitive appeal of the Bayesian approach, we believe there is a good case for a more comprehensive assessment of Bayesian methods and software, building on the foundations laid by the present exercise.

References & Bibliography

Adriaenssens and Ellis (2006) *Uncertainty estimation for monitoring results by the WFD biological classification tools*. Science Report.

Hamill and Ellis (2003) *Approaches to classifying surface water bodies for the Water Framework Directive.* R&D Technical Report EA6285

Kelly, M.G., Juggins, S., Bennion, H., Burgess, A., Yallop, M., Hirst, H., King, L., Jamieson, J., Guthrie, R. and Rippey, B. (2007) *Use of diatoms for evaluating ecological status in UK freshwaters.* Science Report.

Lunn, D.J., Thomas, A., Best, N. and Spiegelhalter, D. (2000) WinBUGS -- a Bayesian modelling framework: concepts, structure and extensibility. *Statistics and Computing*, **10**:325--337.

Glossary of terms

Bayesian statistics

An alternative method of data analysis to the more traditional ‘frequentist’ statistics. Most types of data analysis (e.g. estimating a mean, regression analysis, ANOVA) can be undertaken using frequentist or Bayesian statistics.

Bayes’ Theorem

A simple equation for conditional probabilities, showing how to calculate the probability of [A given B] from the probability of [B given A]. This is the basis for Bayesian statistics, where it is used to calculate the probability of a parameter given a data set.

Doodle

A flow diagram used to represent a statistical model in the WinBUGS software.

Frequentist statistics

The most common statistical basis of data analysis, as implemented with spreadsheets such as Excel, or statistics packages such as Minitab.

Likelihood function

Function describing how the probability of observing a particular data set depends on the values of the underlying parameters of a statistical model. The likelihood function is central to both frequentist and Bayesian statistics.

Parameter

A constant in a statistical model, such as a mean or a variance.

Posterior probability distribution (‘Posterior’)

A probability distribution describing the belief in possible values for a parameter. This is the output from a Bayesian data analysis.

Precision

In frequentist statistics, precision mans the half-width of the confidence interval, whereas in Bayesian statistics, it means 1/variance.

Prior probability distribution (‘Prior’)

A probability distribution describing the belief about the possible values for a parameter, prior to analysing any data. An essential component of Bayesian analysis.

Statistical model

A formal probabilistic description of the relationship between observed data and unobservable parameters. A statistical model lies behind most types of data analysis (estimating a mean, regression analysis, ANOVA etc).

List of abbreviations

BUGS	Bayesian inference Using Gibbs Sampling (software)
CAVE	Combines Appropriate Variance Estimates (spreadsheet tool)
CI	Confidence interval
CofC	Confidence of Class
DARES	Diatoms for Assessing River Ecological Status (WFD tool)
EQR	Ecological quality ratio
PofC	Probability of class
PoM	Programme of Measures.
TDI	Trophic Diatom Index.
WB	Water body
WFD	Water Framework Directive
WinBUGS	The Windows version of BUGS, a Bayesian data analysis package.

Annex A Demonstrating Confidence of Class

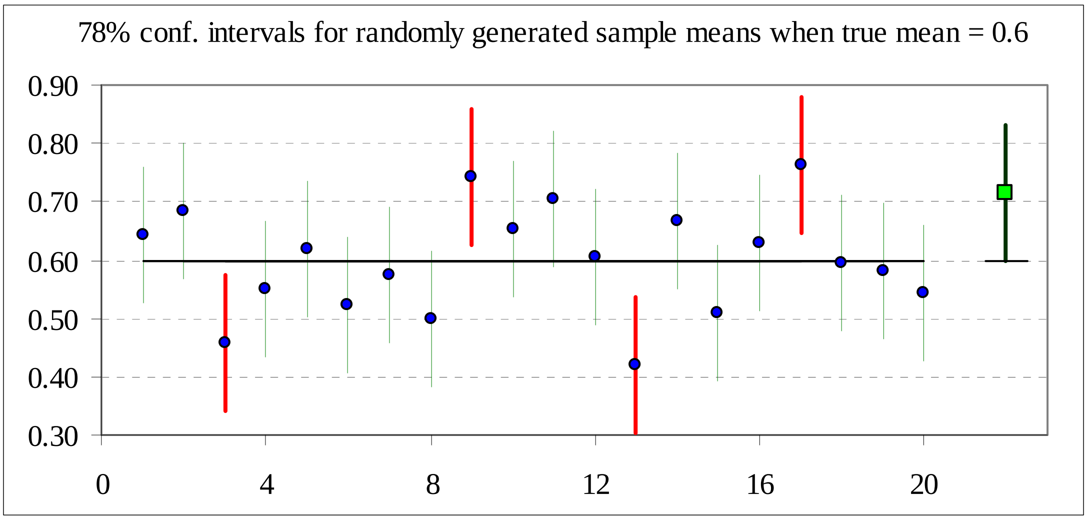
A.1 Setting the scene

Suppose for simplicity that we are interested in just the two EQR class limits defining the ‘Good’ class, namely L = 0.60 and U = 0.80. On the basis of some agreed monitoring programme we obtain a mean EQR of 0.717. We know that the standard error of the mean is 0.095 and that this applies uniformly across the EQR scale from L to U. (That is another simplifying assumption not critical to the argument.) Given this information, how confident can we be that the site is ‘Good’?

A.2 Lower confidence limit

First, consider the situation where the true mean is at L. If we imagine a succession of 20 randomly sampled EQR values, they would be scattered around the true mean as illustrated in Figure 13. The figure also shows the corresponding 78% confidence intervals (CIs) for the true mean. (The reason for the choice of 78% will emerge shortly...)

Figure 13 Randomly generated sample means when true mean = 0.6



We expect about 16 of the intervals (= 78% of 20) to straddle the true mean - and that is what we see here. The other four intervals (shown in red) fail to do so, as they lie either wholly above or wholly below the true mean. In this example, the half-width of each CI is 0.117. So, given any one sample mean, we can be 78% confident that the true mean lies within the interval {sample mean ±0.117}. Furthermore, if we are interested solely in the lower end of the confidence interval, we can be 89% confident that the true mean is greater than the lower confidence limit.

Now we focus on our *actual* sample mean of 0.717 (shown at the right-hand end of the figure by the green square). This lies well above the L boundary of 0.60. But how confident are we that the *true* mean is greater than L? Well, from the previous statement, we can be 89% confident that the true mean is greater than {sample mean - 0.117}, namely 0.717 - 0.117 = 0.60. (And that is why we made the unusual choice of 78% for our confidence coefficient: it was to ensure that the lower confidence limit just touched 0.60.)

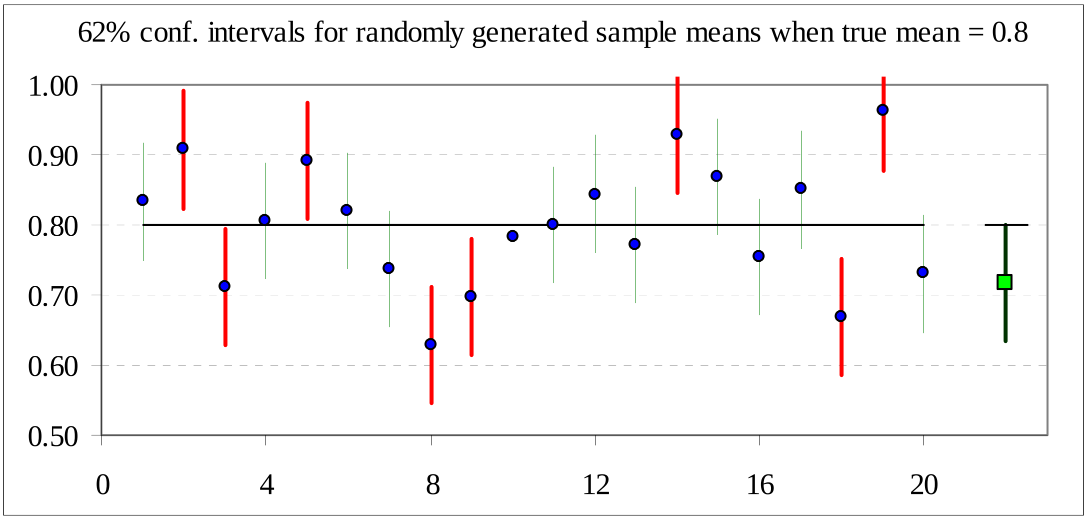
On the basis of the observed EQR estimate, therefore, we can make these statements:

We have 89% confidence that the true EQR is at least as high as 0.60 and
We have 11% confidence that the true EQR is less than 0.60.

A.3 Upper confidence interval

Now consider the situation where the true mean is at U, the upper class limit. Again, a succession of 20 randomly sampled EQR values would be scattered around the true mean as illustrated in Figure 14. The figure also shows the corresponding 62% confidence intervals for the true mean. (And, as before, the figure 62% has been chosen with an eye to the actual EQR estimate).

Figure 14 Randomly generated sample means when true mean = 0.8



We expect about 12 of the intervals (= 62% of 20) to straddle the true mean - and that is what we see here. The other eight intervals (shown in red) fail to do so, as they lie either wholly above or wholly below the true mean. In this example, the half-width of each CI is 0.083. So, given any one sample mean, we can be 62% confident that the true mean lies within the interval {sample mean ±0.083}. Furthermore, if we are interested solely in the upper end of the confidence interval, we can be 81% confident that the true mean is below the upper confidence limit.

Now we return to our *actual* sample mean of 0.717 (again shown by the right-hand green square). This lies well below the U boundary of 0.80. But how confident are we that the *true* mean is lower than U? Well, from the previous statement, we can be 81% confident that the true mean is lower than {sample mean + 0.083}, namely 0.717 + 0.083 = 0.80. We can therefore make these statements:

We have 81% confidence that the true EQR is no higher than 0.80 and
We have 19% confidence that the true EQR is higher than 0.80.

A.4 Confidence of Class

Finally, we can combine these two sets of statements to produce the following CofC values:

Location of true EQR	<L (Moderate)	Between L and U (Good)	>U (High)
Confidence	11%	70%	19%

Annex B The 'Confidence Distribution'

Suppose we have an independent random sample of n = 9 values and we obtain the following summary statistics: mean = 62.8 and standard deviation (s) = 35.1.

We can calculate a 90% confidence interval for the mean in the usual way, viz:
CI = mean ± ts/√n. This produces the values shown below...

Confidence coeff (%)	t	Lower	Upper	Width
90	1.860	41.0	84.6	43.6

There is nothing to stop us repeating the calculation for a whole collection of confidence coefficients. We have done this below in Table 6, for confidence coefficients running from 2% to 99.9%. We see that the confidence intervals start off being very narrow for low confidence levels and grow progressively wider as our desired confidence increases.

Table 6 Confidence intervals for a succession of confidence coefficients

Confidence coeff (%)	t	Lower	Upper	Width	Conf. density
2	0.026	62.5	63.1	0.6	3.30
4	0.052	62.2	63.4	1.2	3.30
6	0.078	61.9	63.7	1.8	3.29
8	0.104	61.6	64.0	2.4	3.28
10	0.130	61.3	64.3	3.0	3.27
20	0.262	59.7	65.9	6.1	3.18
30	0.399	58.1	67.5	9.4	3.02
40	0.546	56.4	69.2	12.8	2.80
50	0.706	54.5	71.1	16.5	2.52
60	0.889	52.4	73.2	20.8	2.16
70	1.108	49.8	75.8	25.9	1.74
80	1.397	46.4	79.2	32.7	1.24
90	**1.860**	**41.0**	**84.6**	**43.6**	**0.66**
92	2.004	39.3	86.3	46.9	0.53
94	2.189	37.2	88.4	51.3	0.40
96	2.449	34.1	91.5	57.3	0.27
98	2.896	28.9	96.7	67.8	0.13
99	3.355	23.5	102.1	78.6	0.06
99.9	5.041	3.8	121.8	118.1	0.00

We have plotted this collection of intervals in Figure 15. The lower and upper ends of each interval are plotted on the 'horizontal' scale and the height of the interval in relation to the 'vertical' axis has been arranged so that, for any confidence interval running from (say) A to B, the area under the curve between A and B is proportional to the corresponding confidence coefficient. (A numerical approximation to this scaling calculation is shown in the final column of Table 6, headed 'Conf. density'.) Thus the (low confidence) narrow intervals near the top of the curve bracket only a thin strip of the area under the curve, whilst the (high confidence) intervals near the base of the curve bracket nearly all the area.

We have coined the term 'Confidence Distribution' to describe this curve, which is specific to the particular data set and required population parameter. It is an unconventional way of looking at statistical confidence and not one that will generally be seen in frequentist textbooks. However, we think it is a helpful device in the context of this exercise as a counterpart to the Bayesian *probability distribution* illustrated in Figure 2.

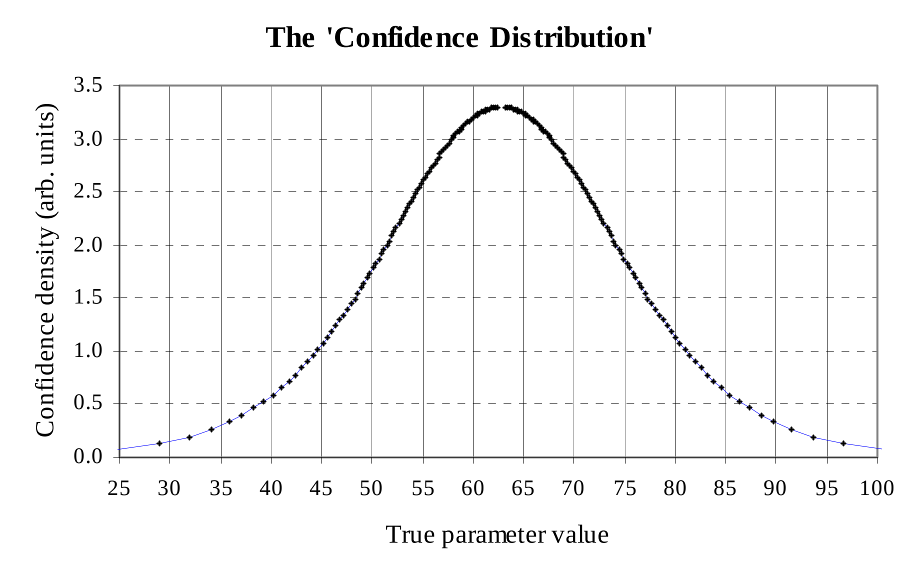


Figure 15 Example of a 'confidence distribution'

Returning to the specific example of the 90% confidence interval introduced at the outset, we see this represented by the dashed purple line in Figure 16 below. As described earlier, the curve has been constructed in such a way that the area within the red shape is 90%; and a similar property holds for all the other confidence intervals in the table.

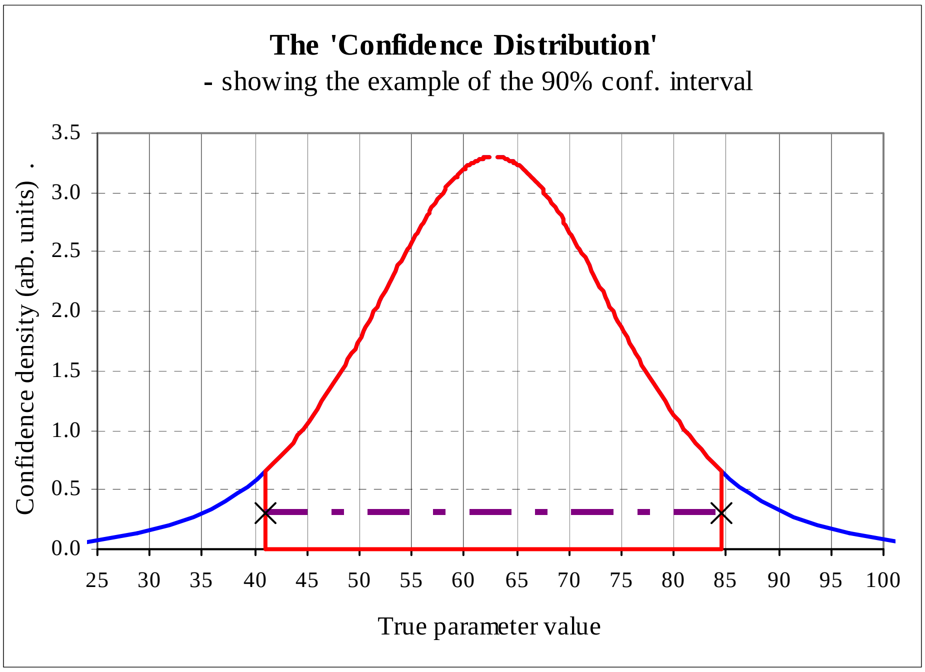


Figure 16 Interpretation of the 'confidence distribution'

We can now disclose the reason for the particular choice of numerical example. The sample mean, 62.8, corresponds to the estimated EQR (0.628) for the hypothetical Site 106 in Table 2. The value of s/√n was chosen to match the expected standard deviation for that EQR value (see Figure 4). Thus, we could determine the confidence that the site was truly in class 'Good' by finding the area under the curve between 52 and 78. (As we have seen in Table 2, the answer is 69.3%.)

The actual method of calculating Confidence of Class (see Section 2.2.2) does not, of course, involve constructing a confidence distribution and calculating the area between any relevant pair of class boundaries. However, the point of the above discussion is to demonstrate that it *could* equivalently be done that way. So to that extent there is a closer parallel, than is at first apparent, between the frequentist's Confidence of Class and the Bayesian's more intuitively appealing Probability of Class.

We are The Environment Agency. It's our job to look after your environment and make it **a better place** – for you, and for future generations.

Your environment is the air you breathe, the water you drink and the ground you walk on. Working with business, Government and society as a whole, we are making your environment cleaner and healthier.

The Environment Agency. Out there, making your environment a better place.

Published by:

Environment Agency
Rio House
Waterside Drive, Aztec West
Almondsbury, Bristol BS32 4UD
Tel: 0870 8506506
Email: enquiries@environment-agency.gov.uk
www.environment-agency.gov.uk

© Environment Agency

All rights reserved. This document may be reproduced with prior permission of the Environment Agency.