Supported by

Environment Agency

NDA

RSK

SAGTA

Professional Guidance: Comparing Soil Contamination Data with a Critical Concentration

CL:AIRE

ISBN: 978-1-905046-35-5

Published by Contaminated Land: Applications in Real Environments (CL:AIRE), Chiltern House, Haddenham Business Centre, Thame Road, Haddenham, Buckinghamshire, HP17 8BY. Web: www.claire.co.uk Email: enquiries@claire.co.uk

Acknowledgements

Soil and Groundwater Technology Association (SAGTA), Nuclear Decommissioning Authority (NDA), Environment Agency and RSK are gratefully acknowledged for their funding of this project. CL:AIRE would also like to thank the steering group members for their reviews of the document.

Author

Nigel Marriott, Marriott Statistical Consulting Limited

Steering Group

Bob Barnes – Environment Agency
Rowan Devlin – East Riding of Yorkshire Council representing YALPAG
Nicola Harries – CL:AIRE
Alex Lee – WSP representing SoBRA
Nigel Marriott – Marriott Statistical Consulting Limited
Ian Teasdale – Sellafield Ltd representing NDA
Hayley Thomas – Shell Global Solutions International B.V. representing SAGTA
Lucy Thomas – RSK
Hannah White – National Grid representing SoBRA
Peter Witherington – RSK

Report Citation

It is recommended citation to this report is made as follows:
CL:AIRE, 2020. Professional Guidance: Comparing Soil Contamination Data with a Critical Concentration. CL:AIRE, Buckinghamshire. ISBN 978-1-905046-35-5.

Disclaimer

This report represents the combined views of the Author and Steering Group and does not necessarily reflect the personal opinions of the Author or of any individual members of the Steering Group. The Author, Steering Group, and Publisher will not be responsible for any loss, however arising, from the use of, or reliance on, the information contained in this document, nor do they assume responsibility or liability for errors or omissions in this publication. Readers are advised to use the information contained herein purely as a guide and to take appropriate professional advice where necessary.

Version Control Sheet

Version number	Version Date	Description of Changes
1	30 September 2020	
1.1	19 February 2021	p3. Changed “simple or stratified random sampling” to “simple random, stratified random or stratified systematic sampling”. p28. Changed “section B4” to “Appendix B4” for consistency. p33. Changed “section B2” to “Appendix B2” for consistency. p34. Changed reference to “Figure B6” to “Table B1”

Background to this guidance

In 2008 CL:AIRE and the Chartered Institute of Environmental Health, supported by the Soil and Groundwater Technology Association (SAGTA) published "Guidance on Comparing Soil Contamination Data with a Critical Concentration". The aim of the 2008 document was to assist land contamination stakeholders (regulators, land owners and occupiers, consultants and others) to apply statistical methods to their data to enable decisions required under the legislative frameworks; either the planning system or Part 2A of the Environmental Protection Act 1990. The document assumed the user had an in-depth knowledge of these regimes and the legal tests they incorporate. The guidance also identified four critical issues that had to be correctly addressed before statistical analysis could be reliably applied to the data obtained from a site assessment:

- Development and refinement of conceptual models.
- Development of appropriate soil sampling strategies.
- Collection and testing of samples for contamination.
- Use of generic and site-specific assessment criteria for risk assessment.

As these critical issues were covered in other guidance, they were kept out of scope, but the document did refer to them wherever appropriate.

The document concluded that the appropriate statistical methods to apply were the Null Hypothesis and Alternative Hypothesis approach set to a defined significance or confidence level. At the time of publishing the 2008 guidance, applying such scientific statistical tests was the widely accepted approach to data analysis which found favour amongst many practitioners by providing a "bright line" for decision making. Although, the guidance highlighted the importance of the conceptual site model (CSM) and understanding the datasets involved, ultimately the application of the scientific test provided a definitive yes/no answer (or bright line) to the question posed.

Over recent years, there has been much international debate over the validity of applying scientific tests based on comparing a p-value to a pre-determined significance level which was formally discussed and summarised by a statement made by the American Statistical Association (ASA) in 2016 (Wasserstein and Lazar, 2016). Although the statement was relating to the application of statistics in scientific research, it is no less relevant to making decisions regarding land contamination. The statement made several points but most importantly for land contamination decisions:

"Practices that reduce data analysis or scientific inference to mechanical "bright-line" rules (such as "$p < 0.05$") for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making. A conclusion does not immediately become "true" on one side of the divide and "false" on the other. Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis. Pragmatic considerations often require binary, "yes-no" decisions, but this does not mean that p-values alone can ensure that a decision is correct or incorrect. The widespread use of "statistical significance" (generally interpreted as "$p \leq 0.05$") as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process."

In its conclusion, the ASA statement made the following recommendation:

"Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning."

Experience in the UK of applying the statistical tests proposed in the previous guidance confirms the points made by the ASA by showing that the "bright-line" approach can lead to over simplistic and erroneous conclusions when the datasets are not fully understood.

This 2020 revision of the guidance addresses the problem of potentially erroneous conclusions by dropping the reliance on a single scientific test and emphasises the importance of a comprehensive understanding of the datasets in the context of the CSM. This will enhance decision making by all stakeholders and promote transparent, reasonable, and justified decisions in regard of land contamination. Using two-way confidence intervals and graphical summaries, the guidance goes on to lead the assessor to question if there is sufficient statistical expertise available in the team to evaluate the data and establish if the dataset is adequate to answer the question posed for the site. Only once assessors are confident that their datasets adequately capture the characteristics of the site being sampled, can a reliable answer be given to the planning or Part 2A questions being asked about the site.

The "scientific reasoning" called for by the ASA, elevates even more the importance of a thorough understanding of the CSM and the other three critical issues identified in the previous guidance and practitioners should not embark upon any analysis until these issues have been fully and competently addressed. However, as with the previous guidance, this document does not provide detailed guidance on them which can, and must, be obtained elsewhere.

As this guidance adopts an entirely different approach to the previous guidance, it completely supersedes it.

Should the reader wish to have further justification for the approach adopted by this guidance, further commentary is provided in Appendix A1.

Application of this guidance

This document provides advice that could be regarded as essential reading towards understanding the different nuances in the regulatory context between the regimes of planning and Part 2A of the Environmental Protection Act 1990. In choosing the best approach a practitioner must first be well-informed as to the different nuances in regulatory context under which they are operating.

A responsibility and requirement exist for critical thinking. Critical thinking may be defined as the process of independently analysing, synthesising, and evaluating information (Hughes and Lavery, 2014). The critical thinker is diligent in seeking relevant information, reasonable in the selection of criteria, focused in inquiry, and persistent in seeking results which are as precise as the subject and the circumstances of inquiry allow.

A practitioner using this document should seek to acquire a full understanding of the potential issues and limitations of their data, together with those of the methods available for its assessment. This document suggests useful methods to summarise, present and assess soil contamination data. However, the application and choice of statistical method should not be considered as a trivial matter.

The contents of this document cannot be considered as being universally applicable; consequently, a number of prerequisites to its use are presented. It is only by being well-informed, for example by reading and understanding the limitations to those methods presented herein, that the critical thinker may identify and justify the most suitable approach for their own site. In essence, purposeful self-regulatory judgment is required that should be well-informed.

The contents of this document are neither mandatory nor definitive and you may wish to use alternative methods to those set out in this guidance. The document does provide an approach to Comparing Soil Contamination Data with a Critical Concentration; it is not definitive but should be considered as valuable guidance informing the critical thinker as to the limitations and factors that they should be considering during their own such assessments.

Readers are advised that they may find it useful to review and satisfy themselves that the following statements are valid for their dataset before using this guidance:

- Averaging Areas and Averaging Zones(1), as well as the Smallest Area of Concern, have been identified on the basis of the CSM, including the desk study and / or the site walkover.

- The sample locations were chosen using a simple random, stratified random or stratified systematic (e.g. a square, herringbone or triangular grid) sampling pattern, rather than being targeted to locations suspected of being contaminated (e.g. stained soil, dead vegetation, location of the underground storage tank on site plans).

- The sample locations are relatively evenly spread across the area and are not clustered, to avoid giving undue weight to some parts of the site over others in the calculated statistics.

- The analyses do not suggest a hotspot or outlier of contamination that should be treated as a separate zone. This has been established by a histogram AND / OR a named statistical test.

- The sample locations are all at the same depth and taken from one population (i.e. the same material) UNLESS the contaminant concentration is independent of these factors for a specified reason.

- Where an Averaging Zone encompasses several Averaging Areas, analyses do not show a spatial trend or other spatial pattern across that zone. This has been established by reporting a spatial analysis such as a post plot of the data AND / OR an experimental variogram.

- The number of samples has been shown to be sufficient for a statistical analysis.

Note (1):

In the context of Human Health Risk Assessment BS ISO 18400-104 'Soil Quality - Sampling' (BSI, 2018) defines the **Averaging Area** as "*that area within which a receptor (e.g. an individual) may be exposed to a source. The size of the relevant Averaging Area is therefore related to the use of the land*".

R&D Technical Report P5-066/TR (Environment Agency, 2000) defines "***Area of interest*** *as that area of soil to which a receptor is exposed or which otherwise contributes to the creation of hazardous conditions – also termed the* ***Averaging Area***".

Averaging Zone is sometimes used to refer to the total area of soil encompassing a similar contaminant. As such an Averaging Zone might contain multiple Averaging Areas. For example, an Averaging Zone of contamination on a site that has been developed with housing could be occupied by numerous gardens, each of which could be analysed as separate Averaging Areas.

For simplicity, the remainder of this document refers only to Zones and the assessor must establish whether it is the Averaging Zone or Averaging Area that is to be analysed to meet the requirements of the investigation.

Contents

Acronyms

ASA	American Statistical Association
ASTM	American Society for Testing and Materials
BS	British Standards
CLT	Central Limit Theorem
CL:AIRE	Contaminated Land: Applications in Real Environments
CSM	Conceptual Site Model
IQR	Inter Quartile Range
ISO	International Organization for Standardization
LCL	Lower Confidence Limit
NDA	Nuclear Decommissioning Authority
SAGTA	Soil and Groundwater Technology Association
SoBRA	Society of Brownfield Risk Assessment
UCL	Upper Confidence Limit
YALPAG	Yorkshire and Lincolnshire Pollution Advisory Group

1. Introduction

This guidance consists of five substantive sections together with three appendices. A brief description of each section is given here to help you know how to get the most of the guidance.

It is essential that you read section 2 ("**How to use this guidance**") first. It lists four prerequisites that must **all** be in place in order for this guidance to be applicable.

Section 3 ("**Datasets used to illustrate this guidance**") introduces three datasets that will be used to illustrate the guidance in sections 4 and 5. Dataset A will be used for a Part 2A scenario, datasets B and C will be used for planning scenarios. This section gives a brief introduction to the datasets but more detail about these datasets can be found in Appendix B. It should be noted that data is a function of the soil type and consequently each of the datasets could be found in either scenario.

All recommended statistical calculations and charts are described in section 4 ("**How to summarise and present your data**"). These can be used for any dataset you have collected provided it satisfies the prerequisites laid out in section 2. All calculations can be done in a spreadsheet such as Microsoft Excel and a list of the spreadsheet functions needed is given in Appendix C. Note, the **Central Limit Theorem** (CLT) is the key to this guidance as discussed in section 5. This is explained in more depth in Appendix A2 and briefly comprises "**The distribution of sample means converge to a normal distribution as the sample size increase**."

Section 5 ("**How to interpret your results**") explains how you can draw conclusions from the calculations and charts you produce using section 4. The main tool used to draw conclusions will be a comparison of a 2-way confidence interval with the critical concentration. However, section 5 makes it clear that such comparisons are not straightforward and you will need to exercise your judgement over a number of other factors that have to be taken into account before you can make your decision.

Section 6 is a checklist of all the outputs you need to pull together in order to describe your decisions and recommendations in a report.

Finally, there are three appendices.

Appendix A expands on the relevance of the ASA statement of 2016 to land contamination problems and explores some of the technical issues behind the statistical approach recommended in section 4.

Appendix B looks at the three example datasets introduced in section 3 in more detail. In particular, this appendix introduces the idea of data type and how this can be a useful technique to help you decide if your sample size is large enough. Data types are also useful for setting expectations on what kind of results you can expect to see and thus can assist you in making decisions about whether your CSM needs further development. Any new user should read Appendix B after section 3 before attempting to apply the remainder of the guidance.

Appendix C contains a list of functions in Microsoft Excel that can be used to perform the calculations explained in section 4.

IMPORTANT NOTE – From now on, this guidance will make frequent reference to mathematical and statistical terms such as **histograms, logarithms, mean, standard deviation, median, quartiles and percentiles**. If you are not familiar with such terms, please read up on these. Appendix C gives a list of some of the functions in Microsoft Excel to calculate these.

2. How to use this guidance

You can only use this guidance if you have devoted time and effort to undertake the four steps listed below. All four steps below must be completed, if you skip any step, this guidance cannot be used.

1. **You have developed a CSM and identified which zone in the site needs sampling and why. Note, this must include clarity on whether you are sampling and using subsequent statistical analysis for the entire source (or averaging zone) or receptor-based averaging area. For more guidance on how to do this, please read these documents**:

    - Standard Guide for Developing Conceptual Site Models for Contaminated Sites, ASTM E1689 – 95, 2014.

    - Soil Quality – Conceptual Site Models for Potentially Contaminated Sites, BS EN ISO 21365:2020.

2. **You have designed your sampling strategy using statistical principles of simple random, stratified random or stratified systematic sampling in order to measure the zone identified by your CSM. For more guidance on how to do this, please read these documents:**

    - Secondary Model Procedure for the Development of Appropriate Soil Sampling Strategies for Land Contamination Development, R & D Technical Report P5-066/TR, Environment Agency, 2000.

    - Guidance on the Classification and Assessment of Waste (1st Edition v1.1). Technical Guidance WM3, 2018. Environment Agency, Natural Resources Wales, Scottish Environment Protection Agency and Northern Ireland Environment Agency. Appendix D within WM3 gives a very good description of sample design and sample size calculations.

    - Soil Quality – Sampling, BS ISO 18400-104, 2018.

3. **You have measured and recorded each sample location separately without compositing samples. If you need to analyse composite samples, please do not use this guidance.**

4. **You have checked and reviewed the data collected from your sample, resolved any issues and inconsistencies, and made a decision on how to deal with non-detects.**

SUMMARY BOX

1. There must be a CSM that identifies which zones you are sampling and why.
2. The sampling strategy must use simple or stratified random sampling to measure the zone(s) identified in the CSM.
3. Samples should not be composite samples.
4. A quality check has been made between the analysis and field data to identify any anomalies and you have established a method for dealing with non-detects.

If the above are not adhered to, decision making based on the statistical results may be unreliable.

3. Datasets used to illustrate this guidance

This section briefly describes three datasets A, B and C that will be used to illustrate sections 4 and 5 of this guidance. All three are inspired by real-life datasets, but the contaminants have not been identified. The measurement scales in all three datasets have been reset using 1 mg/kg as the threshold (the critical concentration) against which the sample will be compared. Finally, all three datasets are samples from a single zone as identified by a CSM.

In section 5 of this guidance, each dataset will be referred to by two descriptors as shown below:

A. Part 2A scenario, Symmetric data.
B. Planning scenario, Log-Symmetric data.
C. Planning scenario, Fat-Tailed data (a probability distribution with a tail that looks fatter than usual, also known as heavy-tailed).

The first descriptor is the scenario that the dataset will be applied to; either a Part 2A decision or a planning decision has to be made.

The second descriptor refers to the type of data being analysed. A full explanation of the three types of data can be found in Appendix B but section 5 will also give an overview. Any new user of this guidance should read Appendix B after section 3 and not simply rely on the information contained in section 5.

Please note there is no relationship whatsoever between the type of data being analysed and the type of scenario being considered. All types of data can occur in any scenario.

Each dataset is briefly described in the next three subsections and the data being used is summarised using a chart known as a histogram. The horizontal axis displays the categories in bands of concentrations 0.1 mg/kg wide which are denoted in brackets like this example (0.9 – 1.0). This is a standard mathematical notation which denotes that the category (0.9 – 1.0) includes values equal to the lower end of the range (0.9 in the example) but does not include values equal to the upper end of the range (1.0 in this example). The vertical bars show the percentage of samples (or sample size if you prefer) that fall within each category. The critical concentration of 1 mg/kg is shown as a vertical black line.

In section 4, you will see the same data displayed in alternative formats (dot plot, box plot, spatial plot) and the results will be interpreted in section 5.

3.1 Dataset A - Part 2A Scenario, Symmetric Data

This dataset is an example where 19 samples have been taken and is summarised in Figure 3.1.

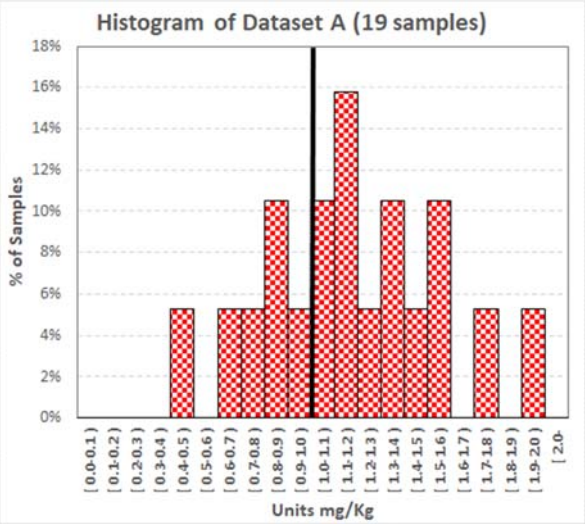

Figure 3.1. Histogram of dataset A.

In section 5.1, this dataset will be used in a Part 2A scenario. However, this dataset could just as easily arise in a planning scenario as well.

3.2 Dataset B - Planning Scenario, Log-Symmetric Data

This dataset is an example where 48 samples have been taken and is summarised in Figure 3.2.

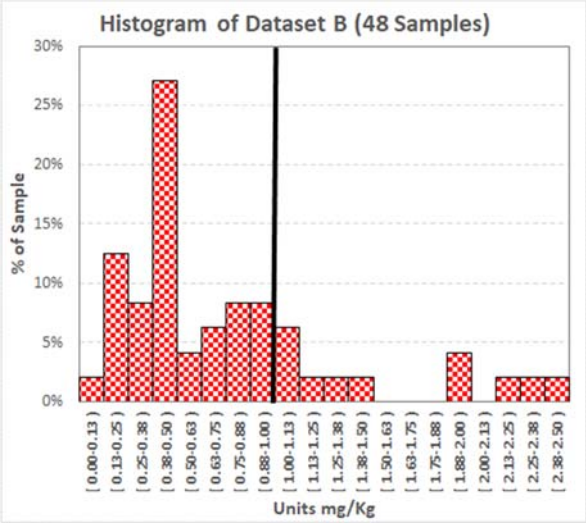

Figure 3.2. Histogram of dataset B.

In section 5.2, this dataset shown will be used in a planning scenario. Again, this dataset can just as easily arise in Part 2A scenarios as well.

3.3 Dataset C - Planning Scenario, Fat-Tailed Data

This dataset consists of 42 samples and is summarised in Figure 3.3.

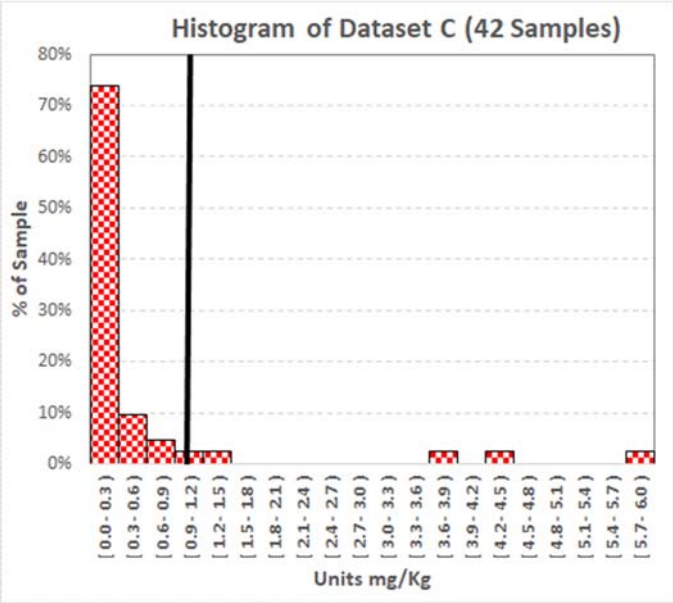

Figure 3.3. Histogram of dataset C.

Dataset C will be used in a planning scenario in section 5.3. Again, it could have been used in a Part 2A scenario instead.

It should be noted that there are other potential data distributions but for the reasons discussed in Appendix B4 in this guidance these are dealt with as symmetric data. Users of this guidance should refer to Appendix B for a more detailed description of the distribution types and how to decide which type bests describes your dataset.

SUMMARY BOX

This section introduces the scenarios (Part 2A and planning) and types of distribution that are commonly found in land contamination studies. The types of distribution are: symmetric, log-symmetric and fat-tailed data. This latter distribution is a probability distribution with a tail that is fatter than usual and is also referred to as a 'heavy-tailed' distribution. A more in-depth description of the types of distribution is given in Appendix B.

4. How to summarise and present your data

This section lists four steps to summarise and present your dataset. This will help with data interpretation and is therefore a crucial part of your statistical analysis. Each step will be illustrated with one of the three example datasets introduced in section 3 and described in depth in Appendix B. This section will give an overview of what the purpose of each step is but the detailed interpretation of the three example datasets will be given in section 5.

The four steps described in this section are:

1. Chart your data with a histogram.
2. Summarise your data with a box plot.
3. Calculate confidence intervals for the mean concentration.
4. Detect spatial patterns with spatial plots.

What is described in this section is in no way an exhaustive list of methods that can be used to present and summarise your data. Many other methods exist but the four step process can be regarded as the minimum needed to interpret your data and make a decision when comparing your data to a critical concentration.

4.1 Chart your data with a histogram (dot plot format)

In any statistical analysis, the first step is to plot the data on a chart and the histogram is a common choice. Appendix B presented histograms for the three example datasets A, B and C which were all presented in the more common format with vertical bars and the categories on the horizontal axis. The histogram in Figure 4.1 for dataset A contains the same data as plotted in Figure B1 but this time the chart has been turned on its side with the categories on the vertical axis and the data presented in a format known as a dot plot. This aids interpretation and further details are provided when the box plot is introduced in section 4.2.

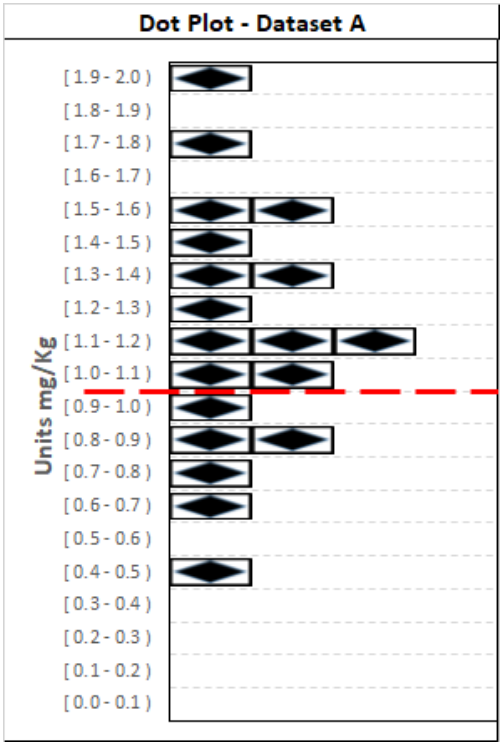

Figure 4.1. Histogram (dot plot format) for dataset A.

The dashed red line represents the critical concentration of 1 mg/kg. In a dot plot format, each sample is shown as a diamond (or some other shape) on the histograms. Given that land surveys sometimes use small sample sizes, this format for displaying the histogram has some advantages. For example, it is very easy to see that of the 19 samples in this dataset, 6 lie below the critical concentration and 13 lie above this level.

Note that there is no horizontal scale on the chart as it is a simple matter to count the number of samples within each category and interpretation will be around the relative importance of each category so even if many more samples were present this should not cause concern. Of course, if the reader wished they could put a note adjacent to each category label with the number of samples or add a horizontal scale.

4.2 Summarise your data with a box plot

Histograms and dot plots are a good starting point for displaying raw data, but they are not the only option. They can be somewhat difficult to use when trying to summarise the data in the form of statistics such as the mean, median, etc. The second type of chart that statisticians like to use to visualise data is the box plot (also known as a box and whisker plot) which is shown in Figure 4.2. This keeps the dot plot from Figure 4.1 on the left-hand side but adds the box plot on the right-hand side.

Box plots are conventionally displayed vertically which is the reason why the histogram was plotted on its side in dot plot format in Figure 4.1 since now both the dot plot and the box plot have the same vertical scale which aids interpretation. Note the critical concentration is now shown as a dashed black line on the box plot only but since the dot plot and box plot have the same vertical scales, it is easy to work out where this line would appear on the dot plot.

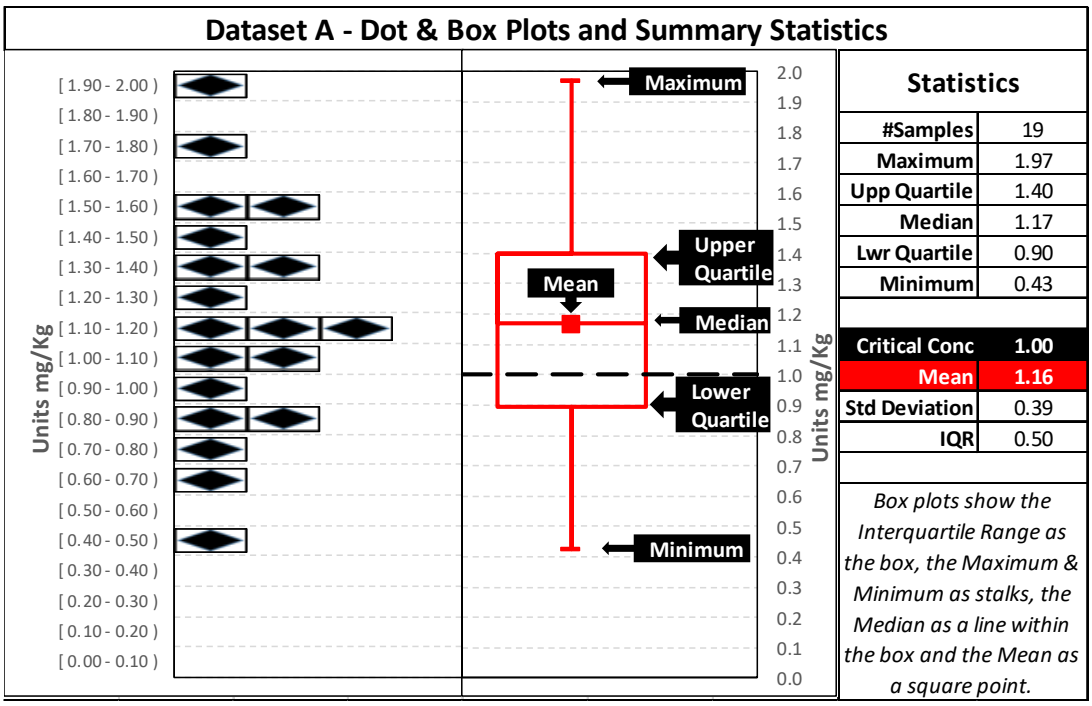

Figure 4.2. Dot & box plot and summary statistics for dataset A.

By plotting the dot plot and box plot side by side, you can get a deeper look at your data but what does a box plot show?

A box plot consists of a **box** and two **stalks** (also known as stems or whiskers) above and below the box. The box contains a horizontal line and a point (a red square block in this example). Together, these features are showing the summary statistics listed on the right-hand side of Figure 4.2 which are respectively, from the top to the bottom:

- The top of the top stalk is the **Maximum** value (1.97 mg/kg).
- The top of the box is the **Upper Quartile** value (1.40 mg/kg).
- The line splitting the box is the **Median** value (1.17 mg/kg).
- The red square is the **Mean** (or **Average**) value (1.16 mg/kg).
- The bottom of the box is the **Lower Quartile** (0.90 mg/kg).
- The bottom of the bottom stalk in the **Minimum** value (0.43 mg/kg).

So, a box plot allows you to see some of the key statistics of the data shown in the dot plot. By plotting the dot plot and box plot next to each other, it is possible to see if any individual sample is particularly influential in any statistic represented in the box plot.

It should be pointed out that box plots come in many forms. Many define the box as stated here but the stalks are sometimes different values such as 5th and 95th percentiles rather than the maximum or minimum. You should be aware of these differences when reading about box plots in statistical textbooks or using different statistical software.

Another name for the box part of the plot is the **Inter-Quartile Range** (IQR) which measures the difference between the upper quartile and lower quartile. In Figure 4.2, the IQR turns out to be 0.50 mg/kg. The IQR can be regarded as a crude measure of how much variation there is in the dataset and is sometimes used as an alternative statistic to the standard deviation.

By definition, half of the dataset lies within the IQR range (i.e. the box) and half lies outside this range as represented by the stalks. Each stalk accounts for a quarter of the dataset and the box is split into two parts by the median, which means each half of the box also accounts for a quarter of the dataset. Therefore, a basic box plot as shown in Figure 4.2 is a simple way of splitting your dataset into four parts, each part corresponding to a one quarter of the sample.

4.3 Calculate confidence intervals for the mean concentration

Figure 4.3 shows the dot plot and box plot for dataset B. It shows that over 75% of the samples lie below the critical concentration (since the upper quartile is less than this value) and the sample mean is also less than the critical concentration. However, Figure 4.3 is based on a sample size of 48 and the question that arises at this point is "*how confident can you be that your estimate of the mean concentration based on your sample is robust enough to draw conclusions?"*

To help answer this question, a measure of confidence is needed, and this can be done by calculating **Confidence Intervals**. These are shown in Figure 4.3 in the box plot as two vertical lines topped and tailed with arrowheads. Two confidence intervals are plotted which are:

- 80% confidence interval (shortest line, dashed). Range is 0.64 to 0.86 mg/kg.
- 95% confidence interval (longest line, solid). Range is 0.58 to 0.91 mg/kg.

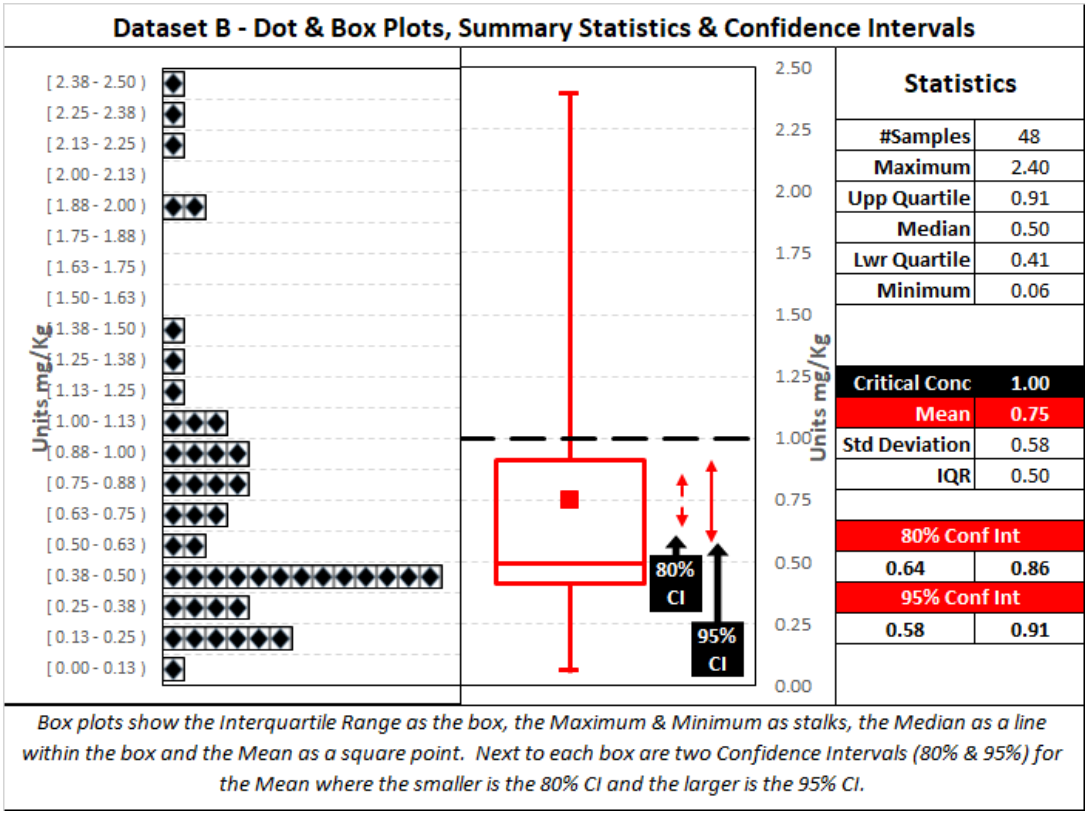

Figure 4.3. Dot & box plot and confidence intervals for dataset B.

What is a confidence interval? A confidence interval consists of two sets of numbers; an **Interval** i.e. a range of values 0.64 to 0.86 mg/kg, plus a pre-determined level of **Confidence** e.g. 80%. A common interpretation of a confidence interval is that "there is an 80% chance that the true mean concentration level of the contaminant in this sampling

zone lies between 0.64 and 0.86 mg/kg". Strictly speaking, the statistical definition of a confidence interval is slightly different which you can find out if you read a statistics textbook (and Appendix A2) but for the purposes of this guidance, this definition is satisfactory.

Formally, the confidence intervals shown in Figure 4.3 are 2-way confidence intervals. This means that if there is an 80% chance of the true mean concentration lying in this interval, there is also a 20% chance it lies outside this interval. Normally, this 20% chance would be split into two equal chunks (hence 2-way) i.e. there is a 10% chance that the true mean is below the lower value of the confidence interval (denoted as LCL) and a 10% chance the true mean is above the upper value of the confidence interval (denoted as UCL). For reasons that are explained in Appendix A2, this is the case for some data types but for other data types, the split is more like 2% chance of being below the LCL and 18% chance of being above the UCL.

The calculation of a confidence interval is a simple matter. The equations for the LCL and UCL are:

- LCL = Sample Mean – T x Standard Error
- UCL = Sample Mean + T x Standard Error

Where:

- Standard Error = sample Standard Deviation / square root of Sample Size
- T = T statistic as derived from the Student T distribution based on the desired confidence level C% and the number of degrees of freedom = sample size – 1.

The T statistics are easily calculated in a spreadsheet (using the TINV() function in Microsoft Excel) or can be looked up from a set of statistical tables. Figure 4.4 shows how the value of T varies given various confidence levels and sample sizes. The confidence levels shown are 80%, 90%, 95% and 99% which tend to be the most common values used but you are free to use other values if you wish.

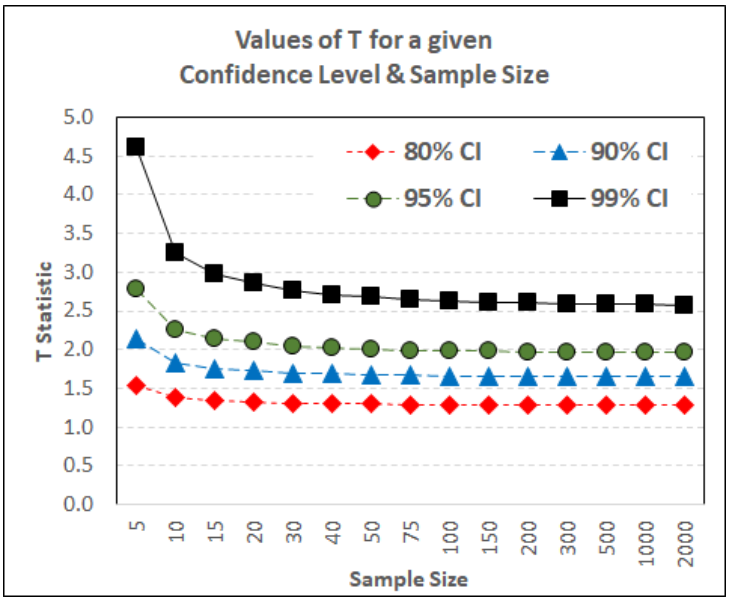

Figure 4.4. T statistics for various sample sizes and confidence levels.

As the sample size increases the value of T stabilises at 1.3 for 80% CI, 1.6 for 90% CI, 2.0 for 95% CI and 2.6 for 99% CI. For smaller sample sizes of 20 or less then T will be notably larger.

Which confidence level should you use? There is no right answer here but it is recommended that at least two confidence levels are used such as 90% and 99% or 80% and 95% which is what appears in Figure 4.3. The purpose of sampling a zone in the first place is to measure the **mean** concentration as accurately as you can but uncertainty will always remain owing to sample limitations. By using two differing confidence intervals, you can use the smaller confidence interval (e.g. 80% confidence interval) to say "I think the mean is most likely to be inside this range" and the larger confidence interval (e.g. 95% confidence interval) to say "I think the mean is most unlikely to be outside this range". It is up to you to decide which confidence levels correlate to the words likely and unlikely but section 5 of this guidance will use 80% and 95% respectively.

4.4 Detect spatial patterns with a spatial plot

The dot plot and box plot give you a good understanding of the underlying distribution of your dataset. However, you should not forget that your data originally came from a 2-dimensional spatial area (or 3-dimensional if you are sampling at different depths). It is possible that the area being sampled is spatially correlated i.e. a measurement at a specific location is much more likely to be similar to a measurement at another location close to it. This would be the case where there has been a spill of a potential mobile contaminant which has migrated away from the location of the spillage.

What is the best way of displaying the data in order to see if a spatial pattern exists?

Firstly, any graphic needs to include a map of the zone being sampled as will have been established in the CSM and the subsequent sampling plan. On that map, you need to show the sampling locations and at each location, put an indicator of what the measurement was. Two obvious options for the indicators are:

- A symbol where the size of the symbol denotes the value of the measurement.
- A number which is the value of the measurement.

This would be known as a spatial plot but this is unlikely immediately to suggest spatial patterns.

A better indication of a spatial distribution would be given by allocating a quarter to each result using the quartile boundaries that can be found from the box plot. The box itself contains the 2nd and 3rd quarters with the median separating these two, the top stalk contains the 4th or top quarter and the bottom stalk contains the 1st or bottom quarter. On the spatial plot each sample location can be annotated with one of four colours, or a number 1 to 4 representing the quarter in which the result lies.

Such a plot should make it easier to spot any spatial patterns as either numbers or colours may start to be clustered in certain areas of the map. Some examples can be seen in Figure 4.5 which uses dataset A. This uses both numbers and colours to denote the quartiles where '1 & green' is the lowest quartile and '4 & red' denote the upper quartile. These maps are to help understand the dataset and establish whether any spatial distribution is what the CSM would have led you to expect. The maps do not show site features as the spatial plot is simply for illustration of the concept. In practice, spatial maps of real life will not be as neat.

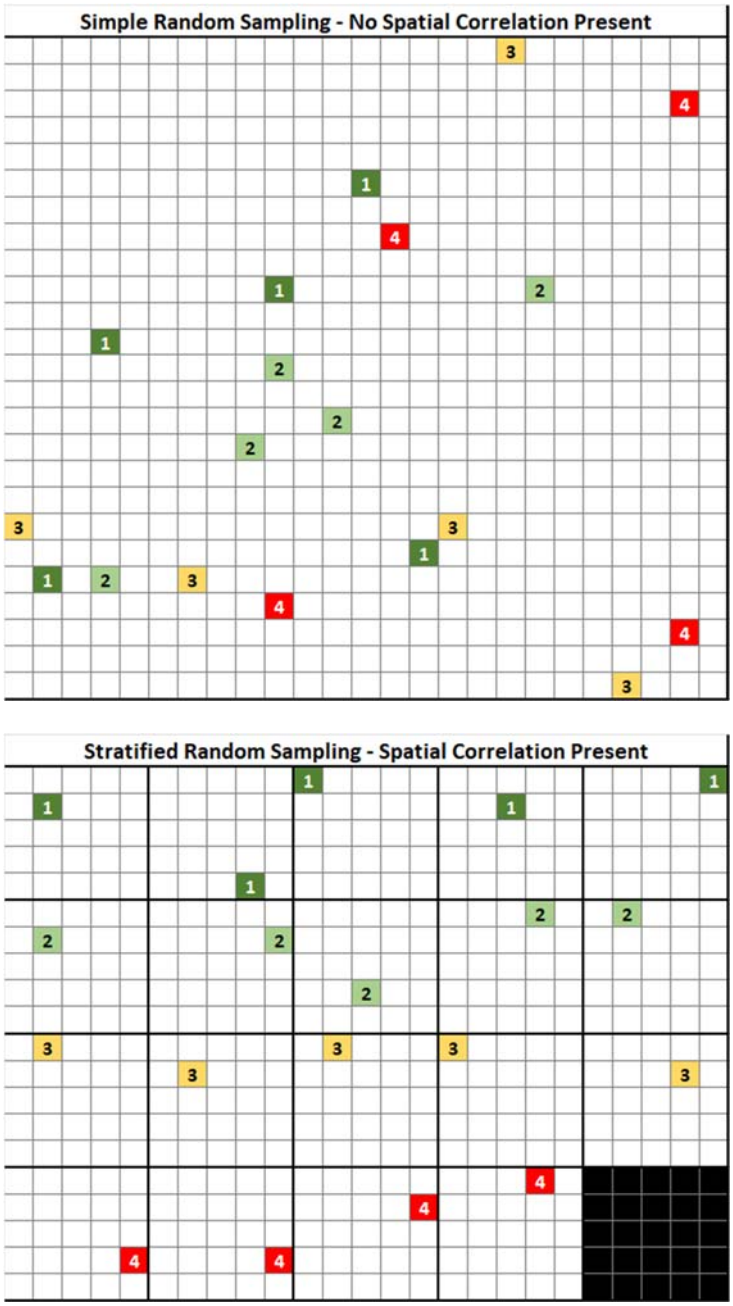

Figure 4.5. Two spatial maps for dataset A.

The first spatial plot assumes the dataset was selected by simple random sampling from a 25x25 grid. No spatial pattern is apparent in this map.

The second spatial map assumes a 20x25 sampling grid was split into 20 5x5 blocks with one block inaccessible (the blacked out one). Within each block one sample was taken using simple random sampling which makes the sampling plan one of stratified random sampling. This time, a clear spatial pattern emerges.

There are many statistical methods for measuring spatial correlations (e.g. variograms, etc) but these are beyond the scope of this guidance. For now, the question to ask is what does it mean for your analysis, if your sample is spatially correlated?

It may mean that you need to revisit your CSM. If you were not expecting any spatial pattern in your dataset, then the presence of one should lead you to reconsider whether your CSM fully understood the site conditions. It may also be necessary to reconsider the sampling plan to establish if it was appropriate in light of the results obtained.

However, if the CSM predicted the spatial correlation demonstrated by the plot, the confidence intervals that you have calculated (as described in section 4.3) are still valid provided your original sampling plan was based on the principles of stratified random sampling only and not simple random sampling. If you used simple random sampling and your spatial plot shows spatial correlation you will not be able to draw conclusions from your data. Under this circumstance you will have to resample the zone using stratified random sampling, before any conclusion over spatial correlation can be reached.

A caveat for stratified random sampling is that your original stratification of the sampling zone should be consistent with the spatial pattern established. For example, if you had originally stratified your sampling zone because you expected the contaminant levels to increase from east to west but when you have collected your data, you find that the contaminant levels increase as you move from west to east, you will need to revisit your CSM before you can take your statistical analysis any further.

SUMMARY BOX

There are many ways to present data in order to make decisions. This guidance focuses on four steps to help comparison of data to a critical concentration and inform decision making. These are:

1. Chart the data with a histogram (dot plot format). The dot plot easily enables the sample size to be viewed along with the critical concentration.
2. Summarise the data with a box plot. This uses the histogram but enables the data to be viewed statistically with the maximum, mean, median, minimum, and upper and lower quartiles.
3. Calculate confidence intervals for the mean concentration. The common interpretation of this is that '*there is a % chance that the true mean concentration level of the contaminant in this sampling zone lies between x and y mg/kg*'. The level of certainty aids decision making.
4. Detect spatial patterns for example using spatial plots. This enables variance in your data to be identified concurrent with your CSM and statistical interpretation that may also inform decision making either by confirming what was expected or by raising issues that require clarification. For example, if a zone of the site could not be investigated but the spatial plot indicates a pattern nearby.

5. How to interpret your results

This section applies the statistical tools introduced in section 4 to the three example datasets described in section 3 and shows you how the results can be interpreted. As will become clear, there are common themes for the interpretation of each type of dataset but there are also differences.

The common themes of interpretation are:

1. Your goal is to get the best answer possible to the question of 'what is the true mean concentration for the contaminant you are measuring?' This is given by the calculated confidence intervals.

2. You will have determined a critical concentration in advance of the analysis. It should be borne in mind that critical concentrations can change because of new knowledge about the risks that are involved.

3. Your conclusions will be based on a comparison of the confidence intervals with the critical concentration. Such a comparison takes advantage of a key tool of statistics known as the Central Limit Theorem.

The differences are:

1. The motivation for your analysis is either for planning purposes or for Part 2A purposes. The former asks if the mean concentration is below the critical concentration, the latter asks if the mean concentration is above the critical concentration. How confident you need to be in your results is likely to be dependent on why you are doing the analysis in the first place.

2. Deciding if you have enough samples to be confident in your interpretation. This decision can be an entire guidance in itself. In this guidance, a shortcut will be used using the idea of data types which is explained in depth in Appendix B but will be briefly covered in this section.

In this section, the interpretation of data for Part 2A purposes will be illustrated using dataset A and the interpretation of data for planning purposes will be illustrated using datasets B and C.

A reminder that the **Central Limit Theorem (CLT)** is the key to this guidance. This is explained in more depth in Appendix A2 but it is worth briefly highlighting what it says here:

"**The distribution of sample means converges to a normal distribution as the sample size increases."**

The confidence intervals calculated as per section 4.3 are based on the explicit assumption that your sample mean comes from a normal distribution regardless of what distribution the individual samples themselves came from. However, the CLT also states that this assumption will only hold if the sample size is sufficiently large.

Therefore, guidance is needed on when this requirement has been met and the interpretation of the three example datasets that follow in this section does this. **However, such guidance cannot and will not be definite and explicit rules such as**

"if the sample size is greater than 25 then good, else bad" do not exist. You must either follow the sampling guidance referred to in section 2 or use the shortcut of data types introduced in this section and explained in depth in Appendix B.

5.1 Dataset A – Part 2A Scenario

For a Part 2A scenario, the question to answer is "*can we confidently say that the mean level of contamination on this land is high relative to some appropriate measure of risk?"* Statistically, this is equivalent to asking *"is there a high probability that the mean concentration of the sampled zone exceeds the critical concentration of 1 mg/kg?"*

Figure 5.1 presents and summarises the 19 samples collected for dataset A.

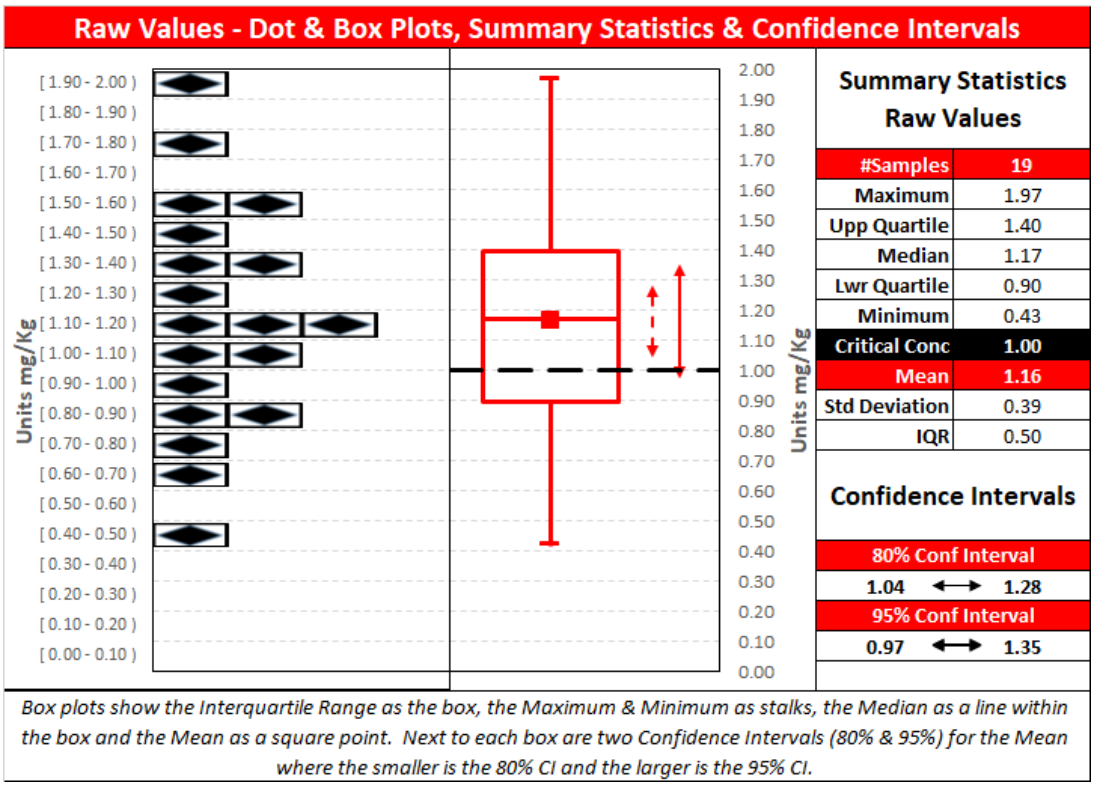

Figure 5.1. A summary of dataset A.

Based on Figure 5.1 "*is there a high probability that the mean concentration of the sampled zone is above 1 mg/kg?"* The sample mean is 1.16 which is above the critical concentration but is there *"a high probability"?* The 80% confidence interval lies completely above the dashed black line, so you can say "*the mean concentration is most likely to be inside this range which happens to be above the critical concentration".*

However, the 95% confidence interval, which measures a range where the mean concentration is unlikely to lie outside of, encompasses the critical concentration of 1 mg/kg so this indicates that there is still a small chance that the true mean concentration lies below the critical concentration.

It is beyond the scope of this guidance to say how certain you need to be (and thus which confidence interval should be used) before you can conclude that the requirements of Part 2A has been met. There is range of possible criteria such as "on balance of probabilities", "beyond reasonable doubt" or "almost certain". The current Statutory Guidance for Part 2A states that decisions should be made "on balance of probabilities"

which justifies a smaller confidence interval such as an 80% confidence interval. On that basis, dataset A will meet the criteria for Part 2A remediation.

In fact, if you are making decisions strictly “on balance of probabilities” in a Part 2A scenario then a case can be made that no confidence intervals are needed at all. All you have to do is see whether the sample mean lies above the critical concentration. If it does, then it is more likely than not that the mean concentration of the sampled zone lies above the critical concentration and you can decide accordingly.

If you are not prepared to make a decision like this, then that indicates that you want to be ‘more certain’ than just mere “balance of probabilities”. The Statutory Guidance makes it clear that it is up to the Local Authority to make this decision but in doing so their judgment must be reasonable. Consequently, it would be prudent for a Local Authority choosing an alternative confidence level to “balance of probabilities” to document their reasons for so doing.

If you are wavering over your decision, then one solution is to take more samples as this will narrow the confidence interval and may do so enough to make a decision clearer but how many samples should you take? If you read a statistics textbook about the CLT, a common rule of thumb used is that the CLT will apply provided your sample size is between 20 and 50.

Such a rule is not definitive but in the case of dataset A, this rule will hold. This is because it is apparent from Figure 5.1 that the type of data being analysed here is **Symmetric.** Symmetric data is one of three types of data explained in Appendix A and the main feature of symmetric data is that the distribution of sample values lying above the median is an approximate mirror image of the distribution of the sample values lying below the median.

Symmetry can show up more clearly on the box plot where the following features will be apparent:

- The height of the box between the upper quartile and median is similar to height of the box between the median and the lower quartile.
- The length of the stalk between the maximum and upper quartile is similar in length to the stalk between the minimum and lower quartile.
- The sample median is close to the sample mean.

For symmetric datasets like dataset A, it can be shown that the CLT holds even with sample sizes less than 20. Even sample sizes of 10 can be sufficient though it should be noted that with such small samples, the confidence intervals could be too wide to make clear decisions.

So, if your CSM and your understanding of the contaminant meant that you expected your data to be symmetric like that of dataset A, then you can proceed with your decision making as described earlier using the confidence intervals.

However, if your CSM and your understanding of the contaminant meant that you expected your data to not be symmetric, but you end up with sample results that are symmetric, then you will not be able to make a decision now. Instead you must revisit your CSM and decide if more samples need to be taken.

5.2 Dataset B - Planning Scenario

For a planning scenario, the question to answer is "*can we confidently say that the mean level of contamination on this land is low relative to some appropriate measure of risk?"* Statistically, this is equivalent to asking "*is there a high probability that the mean concentration of the sampled zone is below the critical concentration of 1 mg/kg?"*

Figure 5.2 presents and summarises the 48 samples collected for dataset B.

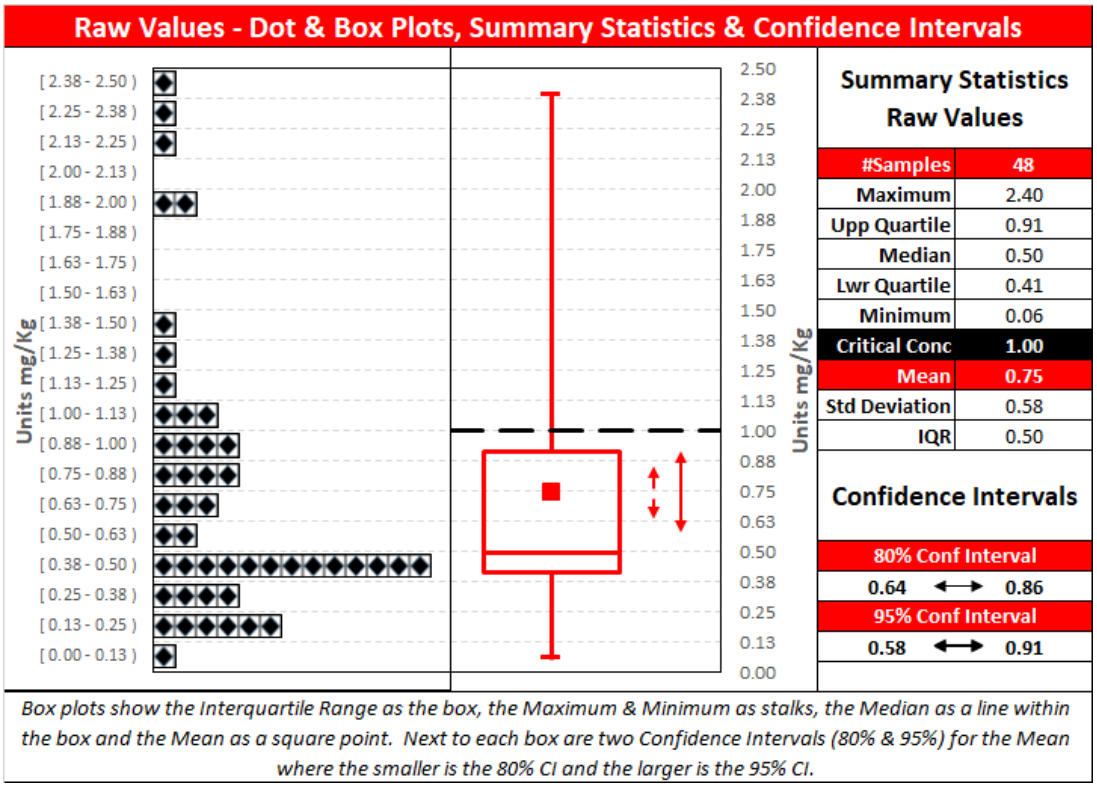

Figure 5.2. A summary of dataset B.

From Figure 5.2, both the 80% and 95% confidence intervals lie below the critical level of 1 mg/kg. The current guidance for planning says that the site has to be "safe" which suggests we should use a large confidence interval such as 95% or higher to make a decision. One could conclude that the requirements of planning have been met since the 95% confidence interval lies below the critical concentration.

However, it is worth looking back at the question posed at the start of this sub-section i.e. "*can we confidently say that the mean level of contamination on this land is LOW relative to some appropriate measure of risk?"* One could argue that whilst the sample mean is less than the critical concentration, it is not "low" relative to it as the sample mean is actually 75% of the critical concentration. Should the critical concentration be revised downwards at a future date, then the confidence intervals may no longer lie below this level.

There is another reason why in this instance you might conclude the mean concentration is not low relative to the critical concentration and that is the veracity of the CLT. It is clear that dataset B is not symmetric, and in this instance, it is an example of **Log-Symmetric** data. Again, Appendix B will explain this data type in more detail, but the key features of log-symmetric data are:

- The distribution of the log values is symmetric.
- The mean is greater than the median but less than the upper quartile.
- The length of the upper stalk from the maximum to the upper quartile is distinctly longer than the length of the lower stalk from the minimum to the lower quartile.
- The height of the upper box from the median to the upper quartile is distinctly longer than the height of the lower box from the median to the lower quartile.
- There are a few sample values at the top of the dot plot that appear at first to be outliers. In Figure 5.2, it looks like there may be 3 to 5 potential outliers. If you have already checked that these are correctly recorded sample values, then it is important to remember with log-symmetric data that apparent outliers are in fact to be expected. They are rare but not unknown and therefore they should not be excluded from your analysis.

It is not difficult to show through computer simulation that when your dataset is log-symmetric, the CLT is not as robust as first thought in two ways (see Appendix A2 for more details):

- First, the apparent confidence levels (80% and 95%) are in fact slightly overstated and the true confidence levels are more like 78% and 93%. The actual discrepancy depends on the sample size with larger discrepancies for small samples.
- Second, when the confidence interval turns out to be wrong i.e. the true mean concentration lies outside the calculated confidence interval, instead of it being equally likely that the calculated confidence interval is too high or too low, what happens about 9 times out of 10 is that the calculated confidence interval will be too low i.e. the true mean lies above the calculated confidence interval. In a planning scenario, such a bias is problematic since it tells us that errors in estimating the true mean concentration are more likely to be underestimates than overestimates. For a Part 2A scenario, this is reversed and therefore this becomes a non-issue.

Suppose you had decided to be “almost certain” before making a decision with the dataset B in Figure 5.2, then these two issues should give you pause for thought. “Almost certain” is probably the same as being at least 95% certain or alternatively less than 1 in 20 chance of being wrong. Currently the 95% confidence interval is below the critical concentration of 1 mg/kg but if you then factor in the two slight issues with the CLT above then you might conclude that your true confidence level is nearer 90% certain or about 1 in 10 chance being wrong. Such confidence might be consistent with “beyond reasonable doubt” but it is not “almost certain”.

On the other hand, if you are prepared to make decisions “on balance of probabilities” then the two issues with the CLT as highlighted would be immaterial with this dataset. However, it should be noted that current guidance for planning purposes does not allow decisions to be made on “balance of probabilities”.

In order to minimise such issues when analysing log-symmetric data, you need to ensure that your sample size is higher than you would have for a symmetric dataset. In section 5.1, mention was made of a rule of thumb for the minimum sample size in order to apply the CLT, namely that this should be between 20 and 50. When analysing log-symmetric data, this rule is still a reasonable one, but the minimum sample size should be closer to 50 as the mean value approaches the upper quartile value and, in some cases, will be higher than 50.

In Figure 5.2, the box plot shows that the sample mean is closer to the upper quartile than the sample median and therefore the minimum sample size, before conclusions using the confidence intervals can be drawn, needs to be nearer 50 rather than 20. This is the case with this dataset since it has 48 samples and so decisions can be made using the calculated 95% confidence interval from Figure 5.2.

If on the other hand, dataset B consisted of only 24 samples but still ended up with the same output as Figure 5.2, then you would not have enough samples to make a decisions using the calculated 95% confidence intervals and instead more samples would have to be taken. Bear in mind that the reason more samples would be needed in this instance is because the 95% confidence interval is close to the critical concentration. If the 95% confidence interval had been 0.1 to 0.4 mg/kg, then clearly this would not be close to the critical concentration and further samples would not be needed.

Finally, the same point made for dataset A is repeated here for dataset B. Dataset B turns out to be of type Log-Symmetric but if your CSM and understanding of the contaminant had led you to expect a dataset that was not log-symmetric, you must revisit your CSM first of all.

5.3 Dataset C – Planning Scenario

Since this is a planning scenario, the question to answer is again "*can we confidently say that the mean level of contamination on this land is low relative to some appropriate measure of risk?*" Again, statistically, this is equivalent to asking "*is there is a high probability that the mean concentration level for the sampled zone is below the critical concentration of 1 mg/kg?*"

Figure 5.3 presents and summarises the 42 samples collected.

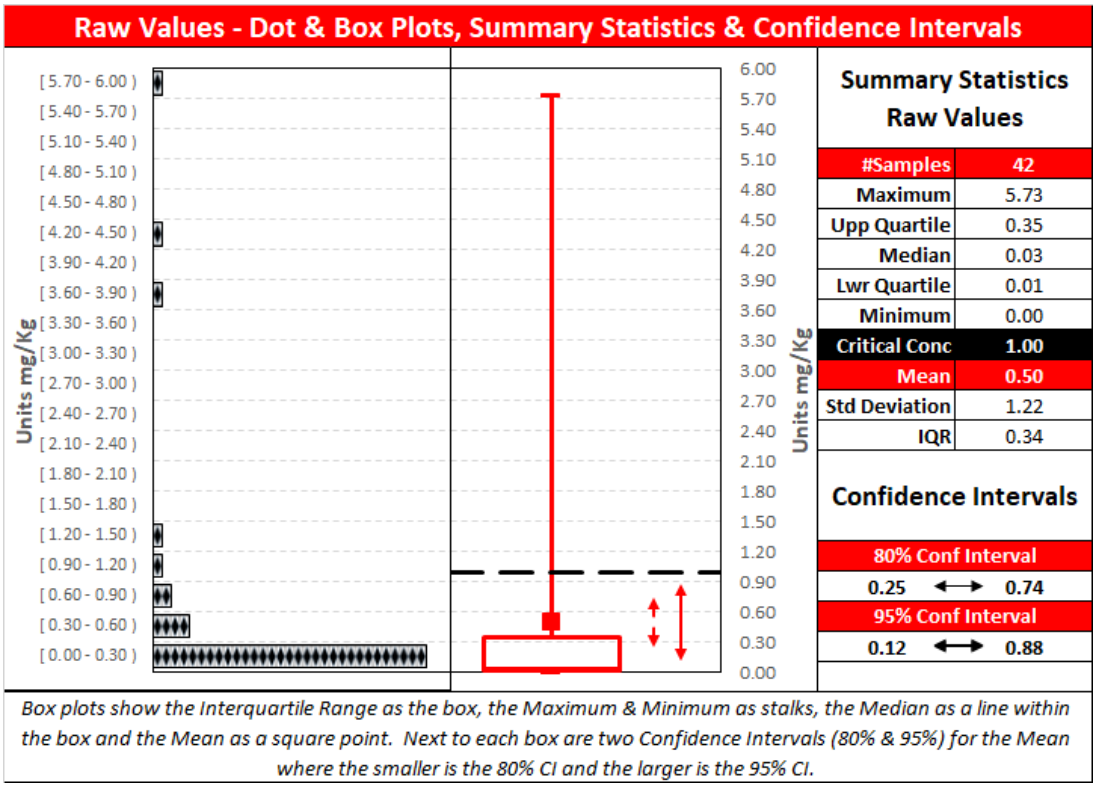

Figure 5.3. A summary of the fat-tailed dataset C.

At first sight, the confidence intervals are below the critical concentration of 1 mg/kg and you could conclude that the planning criteria had been met. Unlike dataset B where the mean value was 0.75 and arguably not "low" relative to the critical concentration, the sample mean of dataset C is 0.5 which is half the critical value. Also, the 95% confidence interval for dataset C is further away from the critical concentration than it was for dataset B.

However, it should be clear that had there been one more extreme value in the sample, the mean concentration (and standard deviation) would immediately be higher and thus the confidence intervals would be wider. Figure 5.4 shows what the results would look like with the addition of a 43rd sample with value of 5.5 mg/kg.

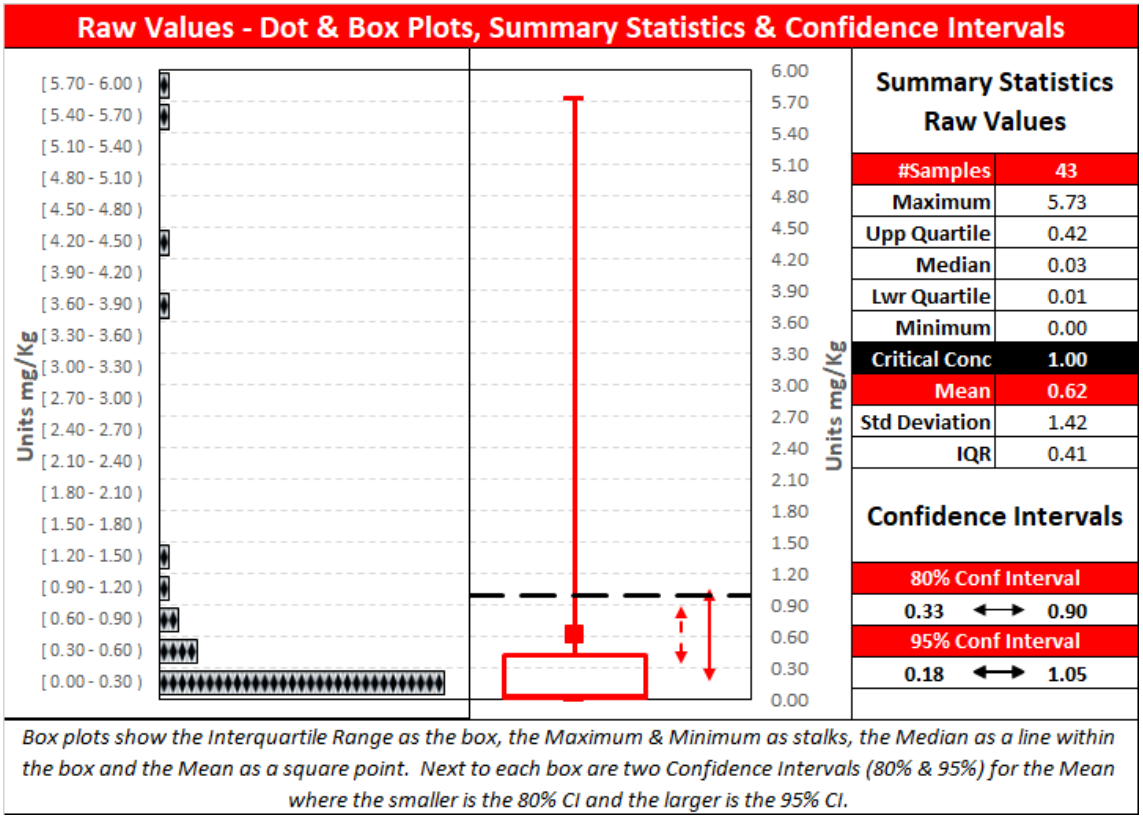

Figure 5.4. Dataset C with additional sample value of 5.5 mg/kg.

Comparing Figures 5.3 and 5.4 shows the sensitivity of your analysis to the presence or absence of extreme values. The sample mean has changed 0.5 to 0.62 and the 95% confidence interval now crosses the critical concentration. This is in fact a classic feature of a third data type known as **Fat-Tailed** data which is what dataset C is an example of. Fat-tailed data is again explained in depth in Appendix B but its key features are:-

- Adding one more sample value with an extreme value has a major effect on your sample mean and confidence intervals.
- The mean lies above the upper quartile.
- The upper stalk from maximum to upper quartile is extremely long relative to the box.

The sensitivity to extreme values is what distinguishes fat-tailed data from log-symmetric data such as dataset B but they share similar issues when it comes to the CLT, only the difficulties with the CLT are greater for fat-tailed data such as dataset C. First, the true

level of confidence you can have in your confidence intervals will be less than the nominal confidence e.g. 95% is more like 90%, etc. Second, when the confidence interval fails to contain the true mean concentration, then the true mean concentration will be found to be higher than the calculated confidence interval. See the latter part of Appendix B2 for more details.

So, to conclude the interpretation of Figure 5.3, it is likely that the planning requirements have been satisfied and that the mean concentration is less than the critical concentration but you cannot be as sure of this as you would be with dataset B.

How can you increase your confidence in your decision when dealing with a fat-tailed data like dataset C? There are two options.

The first obvious option is to take more samples. Table B1 in Appendix B3 shows that previous rule of thumb of taking 20 to 50 samples for datasets A and B is very likely to be too low with fat-tailed data. Appendix B3 describes a rule that can be used to estimate how many samples should be taken which can easily lead to sample sizes over 100. That rule is not the only rule and the sampling guidance referred to in section 2 can be used instead if you prefer. Either way, the end result will be a much larger sample size.

The second option is to accept that you can never reach a burden of proof such as "almost certain" or "beyond reasonable doubt" when making decisions with smaller sample sizes of 20 to 50. Provided your sample contains at least one extreme value (and dataset C in Figure 5.3 does appear to have at least 3 extreme values), it can again be shown through computer simulation that use of the CLT is still acceptable.

If your CSM and understanding of the contaminant has led you to expect a fat-tailed data like dataset C but your sample does not yet contain any extreme values then either you need to revisit your CSM or you need to take more samples. You will not be able to make a decision based on your existing sample.

SUMMARY BOX

This section illustrates data interpretation for the datasets described in section 3.

Always remember that the goal is to get the best answer possible to the questions of '*what is the true mean concentration for the contaminant you are measuring*?'

The decision making will be based on a comparison of the confidence intervals with the critical concentration. This guidance is underpinned by the Central Limit Theorem which briefly assumes '*The distribution of sample means converge to a normal distribution as the sample size increases'*.

Remember to determine whether decisions are required for planning or Part 2A purposes as the questions you need to consider are different. For planning, 'can we confidently say that the mean concentration is below the critical concentration'. For Part 2A, 'can we say on the balance of probabilities that the mean concentration is above the critical concentration'.

If your CSM does not support the distribution recorded by your sample it must be revisited, and you must also consider whether you have sufficient samples to be confident in your decision making. Note there are no set rules in this regard.

6. Making and documenting decisions

In sampling and measuring a zone, your goal is to put yourself in a position where you can make the following statements.

1. “I am confident that my CSM identifies the correct zone to be sampled.”
2. “I am confident that I have used the right sample design to estimate the true mean concentration in my chosen zone.”
3. “I am confident that the sample results are correctly measured and recorded and that I have specified an appropriate rule for non-detects.”
4. “I am confident that my sample size is sufficient to calculate the confidence intervals for the sample mean concentration given that the sample results follow a symmetric/log-symmetric/fat-tailed (or other) distribution.”
5. “I am confident that the true mean concentration lies within the range as calculated by my confidence intervals.”
6. “I am confident that the true mean concentration lies above/below the critical concentration.”

How confident you need to be will depend on the reason why you are sampling a zone in the first place. For example, Part 2A scenarios typically require something to be more likely than not whereas planning scenarios typically require you to be almost sure.

In order to back up your confidence in making these statements, you will need to have carried out the steps described in this guidance (and other documents) and will need to record the following information in any report you write.

1. The regulatory context within which the test is conducted.
2. The rationale for collecting and testing samples including confirmation of the scale of sampling and that unbiased sample data have been used.
3. Details of sampling methods used and justification of the sample size.
4. The process followed to create relevant datasets and the outcome of data quality checks including the identification of any data that have been discarded (with full justification).
5. The methods used to handle non-detects and outliers including identification of any substitute values, any discarded data (with justification) and records of any revised datasets to which statistical tests have been applied.
6. Any expectations as to what kind of dataset should arise.
7. Presentation of any charts and tables used to display the data.
8. A list of the confidence intervals calculated and the reasons why these were considered appropriate.
9. The interpretation of the charts and the confidence intervals.
10. Recommended next step(s).

Appendix A - Approach

A1 Justification of the statistical approach of this guidance

Underlying the entire guidance are the following principles as laid out by the American Statistical Association in March 2016. The sections are aligned with these principles as shown:

"*Good statistical practice, as an essential component of good scientific practice, emphasizes"* …

1. … *principles of good study design and conduct*, …
   - Ensure you have a well-developed CSM.
   - Design an appropriate random sample using statistical sampling methods such as stratified random sampling.
   - Decide what sample size is needed given your objectives and your constraints.
   - Decide on an appropriate non-detect rule for your sample data.
   - Decide on the critical concentration level against which results will be evaluated.

2. *… a variety of numerical and graphical summaries of data,* …
   - Ensure your sample values have been checked, cleaned and investigated and suspicious values are remedied.
   - Plot your data using charts such as spatial plots, histograms, box plots, etc.
   - Calculate summary statistics such as mean, median, standard deviation, quartiles, maxima, minima, etc.
   - Calculate confidence intervals for the mean value.

3. … *understanding of the phenomenon under study*, …
   - Use your CSM and previous experience of similar sites to envisage what kind of results you would normally expect and what kind of results would cause you to go back and revisit your CSM.
   - State what type of data (symmetric, log-symmetric, fat-tailed) you would expect to see in your results.

4. …*interpretation of results in context,* …
   - Decide if your results are consistent with your CSM and expected type of data.
   - Decide if you have taken enough samples to make a decision.
   - Do the results point to a clear and obvious decision or is it more borderline?

5. … *complete reporting* …
   - Have you documented all steps in collecting, summarising and interpreting data?

6. *… and proper logical and quantitative understanding of what data summaries mean. …*
   - Decide if you are in a position to compare the calculated confidence intervals with the critical concentration.
   - If the results are not clear and obvious, what risks are you taking should you end up making the wrong decision? What else do you need to clarify your decision?

7. … *No single index should substitute for scientific reasoning.*
   - You will not find any hard rules in this guidance.
   - There are no 'if this then good, else bad' statements.

- If you are looking for a tick box checklist, you will be disappointed.
- The guidance is written on the assumption that it will be read and used by people with a scientific training who are capable of exercising scientific judgement and who wish to use statistics to SUPPLEMENT their professional judgement, not to REPLACE their professional judgement.

A2 Why does the guidance recommend 2-way confidence intervals and not 1-way?

Confidence intervals can be calculated using the formulae shown in section 4.3 because of a well-known result in statistics called the **Central Limit Theorem (CLT)**. This states that regardless of the underlying distribution of the data (as shown by a histogram), the distribution of possible values for a sample mean will increasingly follow a normal distribution with mean value equal to the population mean and a standard deviation equal to the standard error (as calculated in section 4.3) as the sample size increases. In practice, the T distribution is used since the T and Normal distributions are one and the same for large sample sizes, it is only for small sample sizes they differ.

A 2-way confidence interval calculates an upper and lower bound for the mean concentration whereas a 1-way confidence interval calculates only an upper bound (for planning purposes) or a lower bound (for Part 2A purposes). A 1-way confidence interval might appear to be more useful but this guidance recommends using a 2-way confidence interval for two reasons.

First, it encourages the reader to focus on estimating what the true mean concentration is for the zone being sampled and to ask whether their sampling strategy and CSM has been successful in this endeavour. The second reason is that 2-way confidence intervals are less sensitive to errors that can arise when the CLT is assumed to be true for datasets that are either log-symmetric or fat-tailed. This can be demonstrated through computer simulation.

For example, suppose you were to use the randomisation feature of Microsoft Excel (the RAND() function) to select a random sample of 30 for a population where the contaminant is known to follow a log-normal distribution with mean value of 1.3 mg/kg and a standard deviation of 1.3 mg/kg (on the log scale, this is for a mean of 0 and standard deviation of 0.75 similar to that seen for the log-symmetric dataset in Appendix B2). The mean and standard deviation of this simulated sample can then be calculated along with a 2-way confidence interval. For the purposes of this discussion a 90% confidence interval will be used but the same points can be made with any other confidence interval.

Let's suppose the first simulated sample 90% confidence interval is 1.1 to 1.4 mg/kg. This can be regarded as a "hit" since this interval contains the true mean of 1.3 mg/kg. Suppose you now repeat this, and the second sample's 90% confidence interval is 0.95 to 1.25 mg/kg. This would be a "miss" and more specifically an underestimate since the true mean lies above the sample's confidence interval. Similarly, if your third simulated sample results in a confidence interval of 1.4 to 1.6 mg/kg, this would also be a miss, but it would be an overestimate since the true mean lies below the confidence intervals.

Imagine 2000 simulated samples have been created and the number of hits and misses are counted. If the CLT is perfect, then you would expect 90% of the simulated samples to be hits and 10% to be misses with 5% too low and 5% too high. This is in fact the formal definition of a confidence interval. It was noted in section 4.3 that the definition of

a confidence interval used elsewhere in this guidance is satisfactory but is not strictly correct. The strictly correct definition is the one in this paragraph.

In practice, for our log-normally distributed population being used as an example, you would get an outcome as shown in Figure A1 below, which has the following features:

- The black curve is the log-normal distribution for the population.
- The dashed black-yellow line is the true population mean of 1.3 that a simulated sample is supposed to estimate.
- The combined red and green histogram is the distribution of the 2000 simulated sample means.
- The green histogram is for simulated samples that resulted in a “hit” based on their 90% confidence intervals.
- The red zone on top of the green zone is for the simulated samples that resulted in a “miss” based on their 90% confidence intervals.

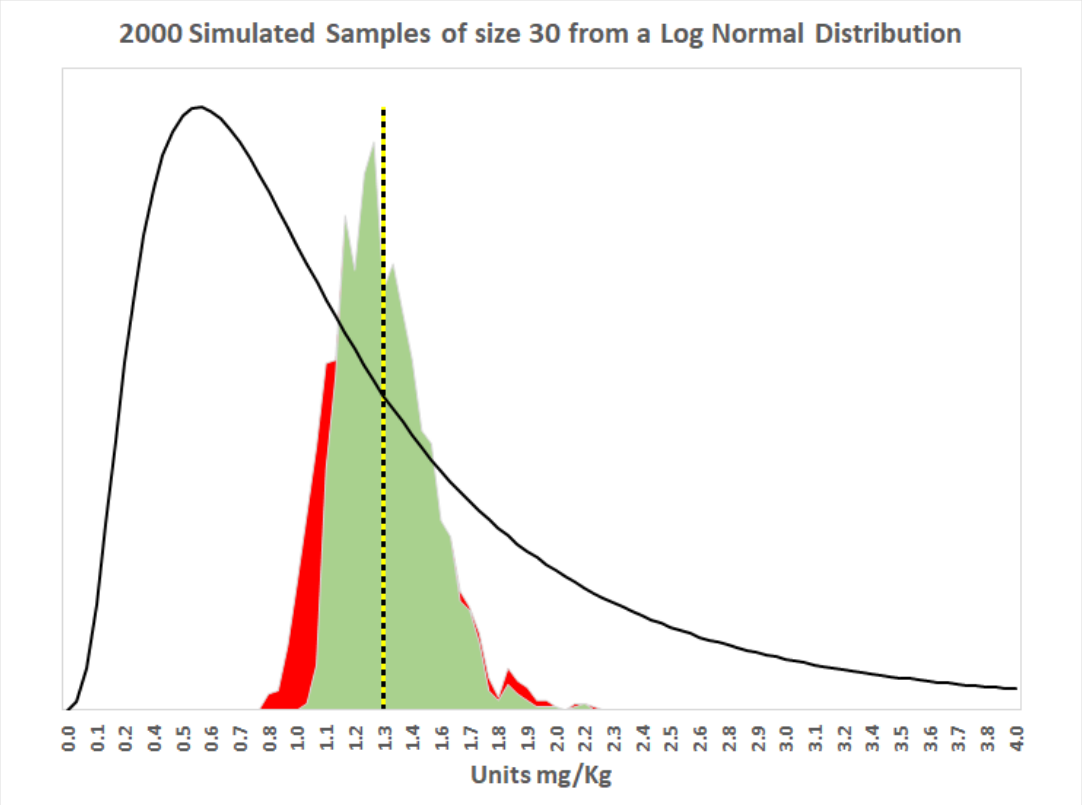

Figure A1. Log-normal distribution and histogram for 2000 simulated samples.

- For this scenario, the true hit rate turned out to be 88% rather than 90%, which is still pretty close. More notable though is that the misses are not balanced. The red zones above the true mean (where the confidence interval is too high) are much smaller than the red zones below the true mean (where the confidence interval is too low). In fact, 11% of simulated samples here are too low and 1% are too high.
- This exercise can be repeated for different confidence levels and sample sizes. Table A1 shows what the actual % of hits and misses are for both 2-way and 1-way confidence intervals. The results are dependent on the underlying distribution but the results are quite clear, 2-way confidence intervals have a

higher hit rate than 1-way confidence intervals and are not far off from the theoretical hit rate as specified by the CLT.

- The results for fat-tailed data are quite similar except that sample sizes have to be higher as demonstrated in Figure B6. In fact, it was this simulation exercise that revealed the rule of thumb 2d stated in Appendix B4, namely that any dataset where the mean value exceeds the upper quartile should be regarded as fat-tailed data. For a sample size of 50 for a log-symmetric distribution where the mean value equals the upper quartile, a 2-way 99% confidence interval will actually have a 91% hit rate which is a considerable divergence from the CLT and hence the need to increase the sample size as specified in Figure B6.

Table A1. Results from the 2000 simulated samples for different confidence intervals and two sample sizes of 30 and 50.

2-Way Confidence Intervals - CLT Accuracy								
Data type	Log Normal				Log Normal			
Conf Interval	80%	90%	95%	99%	80%	90%	95%	99%
Sample Size	30				50			
Outcomes for 2000 simulated samples								
Too High	5%	1%	0%	0%	8%	1%	1%	0%
Hit by CI	78%	88%	92%	96%	78%	88%	93%	97%
Too Low	17%	11%	7%	4%	14%	11%	7%	3%

1-Way Upper Confidence Intervals - CLT Accuracy								
Data	Log Normal				Log Normal			
Conf Interval	80%	90%	95%	99%	80%	90%	95%	99%
Sample Size	30				50			
Outcomes for 2000 simulated samples								
Hit by CI	73%	83%	89%	95%	75%	84%	90%	96%
Too Low	27%	17%	11%	5%	25%	16%	10%	4%

Appendix B – Datasets used to illustrate this guidance

Note: in this appendix histograms are presented in their usual form with the bars vertically, but for the box plots in section 5, the histograms have been rotated by 90 degrees to enable interpretation alongside a box plot.

The purpose of this appendix is to introduce the three types of data that you are likely to encounter in surveys of land contamination. Identifying what type of data you are working with is referred to in sections 3 and 5, particularly in determining how large your sample needs to be to allow a robust interpretation of the results. However, when it comes to the calculations described in section 4, it does not matter what type of data you are working with, the calculations will be the same.

The three types of data are referred to in this guidance by the following names:

1. Symmetric data
2. Log-symmetric data
3. Fat-tailed data

All three types of data are illustrated in this guidance using three different examples inspired by real-life datasets, but the contaminants are not identified. In all three types, the scales have been reset using 1 mg/kg as the threshold (the critical concentration) against which the sample will be compared with. Finally, all three datasets are samples from a single zone as identified by a CSM.

Each dataset is described in more detail in the next three sub-sections. Knowing what kind of dataset you are working with can make it easier to make decisions as demonstrated in section 5 of this guidance. Appendix B4 will help you to decide which example (symmetric, log-symmetric or fat-tailed) is closest to your dataset. Note that Appendix B4 relies on some calculations which are described in section 4 so you will need to be familiar with these before you can decide what type of data you are working with.

B1 Symmetric data

For this example, 19 samples have been taken and summarised using a histogram as shown in Figure B1. The horizontal axis displays the categories (bands of concentrations 0.1 mg/kg wide) and the vertical bars show the % of samples (or number of samples if you prefer) that fall within each category.

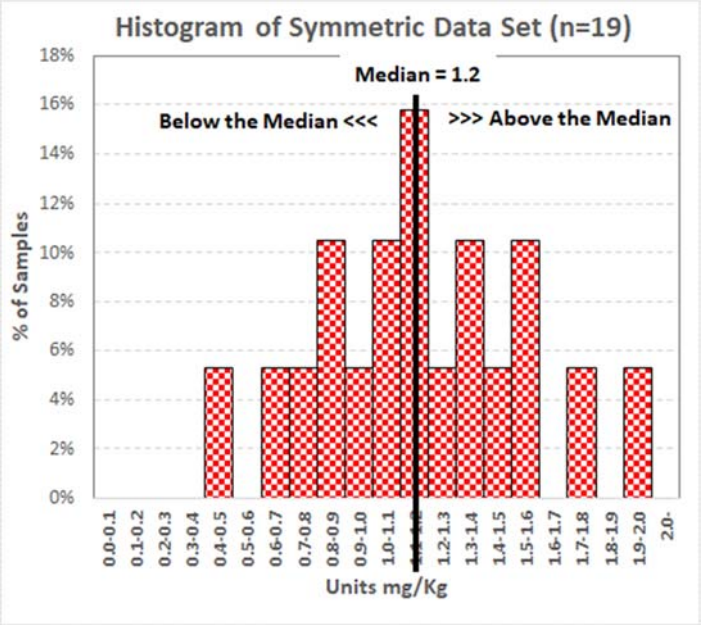

Figure B1. Histogram of the symmetric dataset.

A dataset can be described as **Symmetric** if the shape of the data lying below the median value (1.2 in this example) is a mirror image of the data lying above the median value. This is the case in Figure B1.

The exact shape of the data either side of the median is immaterial and it is not necessary for both sides to be perfect mirror images. A clue that your dataset is probably symmetric is when the mean and median values are close together. In this dataset, the mean is 1.17 mg/kg and the median in 1.16 mg/kg, a difference of less than 1%. This is more than close enough given that the 19 samples vary between 0.4 mg/kg and 2 mg/kg.

Note, there is no statistical definition of “close enough”. One could say “any difference less than 5% is close enough” but you have to take into account the underlying variation in the data as well. You will find the box plot introduced in section 4.2 a very useful tool to help you make this determination so do not worry if you are unclear at this point in time on how to decide when the mean and median are close to each other.

Another helpful indication of whether a dataset is symmetric is when each of the quarters are of similar size. In Figure 4.2, both stalks are roughly equal in length, the box has been split into two roughly equal parts, each of similar length and the mean sits on the median line. Therefore, this dataset is clearly symmetric.

A well-known symmetric dataset is the Normal distribution, also known as the Gaussian distribution or informally as a bell curve. The data in this example is in fact an example of a normal distribution and Figure B2 shows a normal distribution overlaid on the histogram.

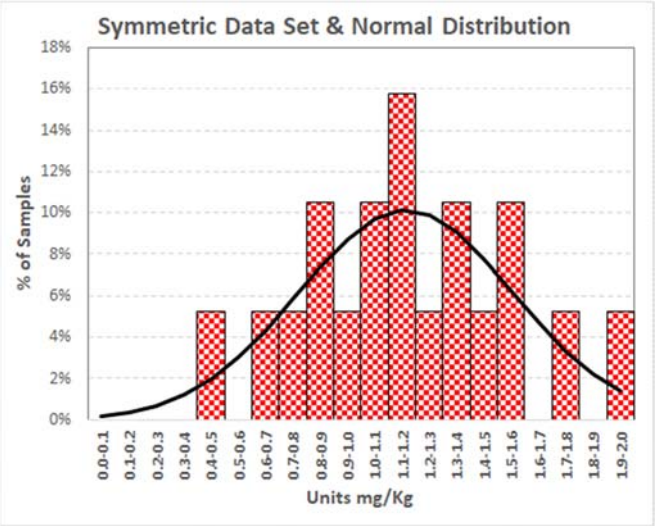

Figure B2. The normal distribution overlaid on the symmetric dataset.

What kind of situations are likely to give rise to a symmetric dataset in real life? The most plausible example is when there is a reasonable amount of contamination in land that has been thoroughly mixed up and deposited such as mine spoil or industrial waste or the contaminant is of a type that can diffuse and migrate through the ground.

B2 Log-Symmetric Data

The log-symmetric dataset is not symmetric, and the box plot of this data shows this very clearly in Figure 4.3. The two stalks have very different lengths, the median splits the box into two unequal chunks and the mean is very different from the median and is closer to the upper quartile. This makes it easy to identify this dataset as log-symmetric and illustrates the rules of thumb introduced in Appendix B4 very well.

To illustrate this type of data, an example where 48 samples have been taken will be used. The 48 values are displayed in Figure B3 along with a black line showing what would be expected if this dataset was log-normally distributed. The log-normal distribution is a common example of a log-symmetric distribution.

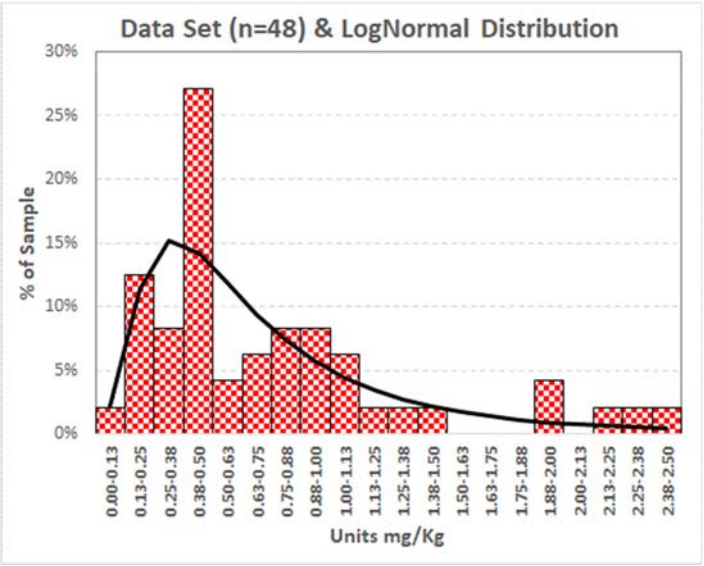

Figure B3. Histogram of log-symmetric dataset.

A log-symmetric dataset is an example of an asymmetric or skewed dataset. Most of the sample values are bunched near the lower end of the scale but as you go up the scale, fewer and fewer values appear.

A dataset is log-symmetrically distributed if the histogram of the log values is a symmetric distribution as defined in Appendix B1. You can calculate the natural logarithm of a value in Microsoft Excel using the LN() function provided the value is greater than zero (you will need to ensure that you have a non-detect rule that gives non-zero values). For the dataset in this example, the histogram of the log-values is shown in Figure B4 and what is shown there is sufficiently symmetric. The black line shown in Figure B4 is what would be expected if the log-values followed a normal distribution. As you can see, the black line and blue bars are roughly the same shape. In such cases, the data is said to be a log-normal distribution.

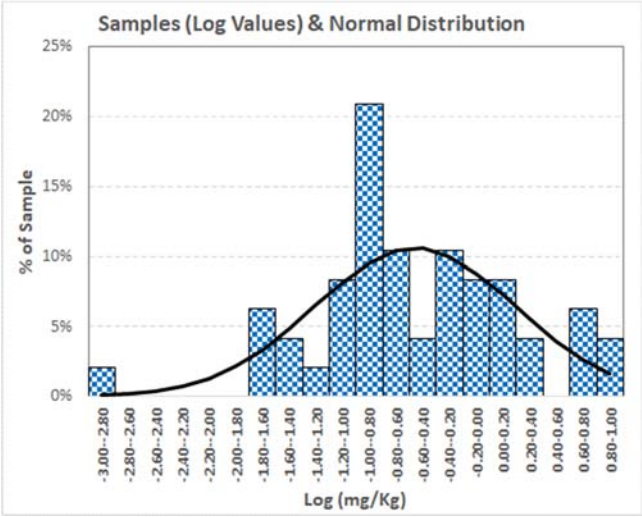

Figure B4. Histogram of log-transformed dataset.

There are other clues that point to your dataset being log-symmetrically distributed. The main one is that the mean value is notably higher than the median value. In this example, the median value is 0.5 mg/kg and the mean value is 0.75 mg/kg which is 50% higher than the median. On the log scale though the median is -0.7 and the mean is -0.56 and as described in section 3 for symmetric data, these values are close enough to regard Figure B4 as symmetric on the log scale and hence log-symmetric on the actual scale. Again, there is no precise definition of “close enough” but a difference between mean and median on the log scale of less than 0.2 is a reasonable rule of thumb.

In section 5.2, the example dataset shown in this section will be used in a planning scenario. This does not mean that log-symmetric datasets cannot occur in a Part 2A scenario as the distribution is only a function of the soil type and not the scenario being applied.

An example of soil with a log-symmetric dataset would be when contamination with a material that does not break down easily has been introduced into a less contaminated matrix and not been particularly well mixed.

B3 Fat-tailed data

The dataset in Figure B5 (consisting of 42 samples) is a classic example of fat-tailed data. At first sight, the data shown in Figure B5 look like log-symmetric data since we have a lot of samples near the minimum value and few samples near the maximum. The mean value of 0.5 mg/kg is considerably higher than the median value of 0.03 mg/kg.

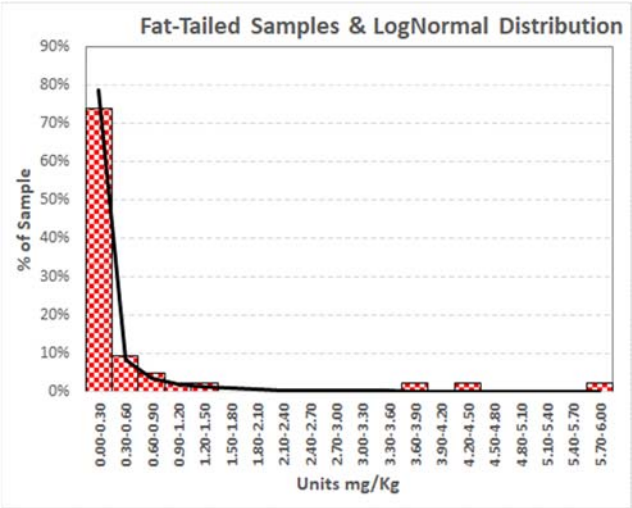

Figure B5. Histogram of fat-tailed dataset.

In fact, the extreme difference between the mean and median values in this case (the mean is almost 20 times the median value) is a clue that the data is not log-symmetric. A log-symmetric dataset (see Appendix B2) is symmetric when the log values are plotted on a histogram and Figure B6 shows that this is not the case for fat-tailed data. This is reinforced by the fact that the mean log value of -2.86 is notably higher than the median log value of -3.43 (recall the suggested rule of thumb in Appendix B2 of a difference between mean and median log values being less than 0.2).

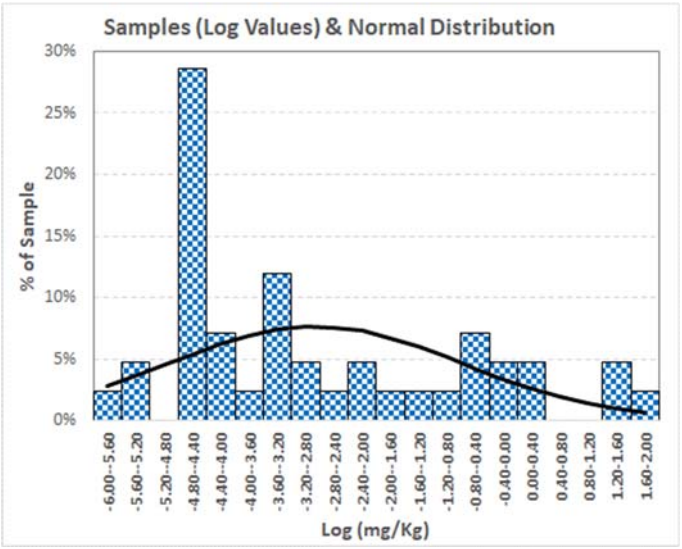

Figure B6. Histogram of log-transformed fat-tailed dataset.

When fat-tailed data is referred to in this guidance, the following terminology will be used.

- **Main body** – refers to the bulk of the data which will typically be bunched close to the minimum value. In our dataset, 39 of the 42 samples could be regarded as the main body based on Figure B5.

- **Extreme values** – refers to the extreme values that differ markedly from the main body. Figure B5 suggests that 3 out of 42 samples are extreme values.

Section 5 of this guidance describes how to interpret the results of your dataset and when working with fat-tailed data a useful statistic is the **expected frequency of extreme values**. This is expressed as 1 in X and so for this dataset, the observed extreme value frequency is 1 in 14 (=3/42). Another way to think of this statistic is to ask how many exposures do you need before you can be expected to be exposed to an extreme value? You need to make an educated guess of what you think the extreme value frequency will be if you are sampling a zone you expect to be fat-tailed since your required sample size will be dependent on this statistic.

The example dataset shown in this section will be used as a planning scenario in section 5.3. However, distribution is a function of the soil type and hence fat-tailed datasets could equally well occur in the Part 2A scenario.

Table B1 overleaf shows that the rules of thumb for the minimum sample size being between 20 and 50 as described in section 5.2 for log-symmetric data is unlikely to be sufficient when dealing with fat-tailed data. The calculations used in Table B1 come from a field of statistics known as **Attribute Sampling** which is a widely used method in many industries and is covered by the British Standards BS6001 series. If you are regularly sampling contaminants known to be fat-tailed then you should familiarise yourself with the statistics described in those standards.

It is not difficult to work out how large your sample needs to be to have confidence that it contains at least one extreme value. Suppose the zone you are sampling has an extreme value frequency of 1 in 10 exposures. This means the probability of a sample not containing an extreme value is 9 out of 10. If you are using random sampling, then the probability of your first two samples not containing extreme values is 0.9 times 0.9 or 0.9 squared. Continuing this thought process, the probability of taking N random samples and none of them having an extreme value will be 0.9^N (0.9 raised to power of N).

More generally, if your extreme value frequency is 1 in X and you are taking N samples, then:

$$\textbf{Probability (no extreme values in your sample)} = P0 = [\,(X-1)/X\,]^{N}$$

If you wanted to be at least 95% certain of having at least one extreme value in your sample, this is the same as saying P0 needs to be less than 5% (or 0.05). By changing this equation to an inequality first and then rearranging, you can get an equation for the minimum sample size with fat-tailed data i.e.

$$P0 < [\,(X-1)/X\,]^{N} \qquad \text{i.e.} \qquad N > \mathrm{Ln}[\,P0\,] \,/\, \mathrm{Ln}[\,(X-1)/X\,]$$

Where Ln() denotes the natural logarithm (LN() function in Microsoft Excel).

Suppose you want P0 to be less than 1% and you are sampling a zone where you expect an extreme value frequency of about 1 in 10, then your minimum sample size will be 44 (= Ln(0.01)/Ln(0.9)). However, if your expected extreme value frequency is 1 in 50,

then the minimum sample size will be 228. Table B1 shows a variety of minimum sample sizes for fat-tailed datasets for differing values of P0 and X.

Table B1. Table of minimum sample sizes for fat-tailed data.

Minimum Sample Sizes for Fat-Tailed Data				
		Probability No Extreme Values in Sample		
		P0 = 10%	P0 = 5%	P0 = 1%
Expected Extreme Value Frequency (1 in X exposures are extreme values)	5	10	23	51
	10	22	49	112
	15	33	76	173
	20	45	102	234
	25	56	129	295
	30	68	155	356
	35	79	182	417
	40	91	208	478
	45	102	235	539
	50	114	261	600
	55	125	288	662
	60	137	314	723
	65	149	341	784
	70	160	367	845
	75	172	394	906
	80	183	420	967
	85	195	447	1028
	90	206	473	1089
	95	218	500	1150
	100	229	526	1211

For the fat-tailed dataset shown in Figure B5, it was observed that the extreme value frequency was 1 in 14. Suppose this is close to what you would have normally expected with this contaminant, then Table B1 shows that your minimum sample size needs to be between 33 and 173 (reading from the 1 in 15 row) depending on how certain you want to be that your sample contains extreme values. The actual sample size of 43 is at the lower end of the range but is probably sufficient.

It is worth pointing out that another way of achieving larger sample sizes is to analyse composited samples instead. This guidance is only applicable to non-composited samples and composited samples is beyond its scope. However, compositing is a way of increasing the statistical sample size without increasing the sample size to be measured in a laboratory which is ultimately the main driver of the cost of sampling.

An example of a soil with a fat-tailed dataset would be when diffuse contamination has been introduced intermittently into a wider spoil body. This could be a site when a building containing asbestos has been demolished without sufficient care and asbestos-containing material has been spread randomly on the site. However, another possibility might be that contamination has been introduced as a result of historic activities, which could have been predicted by greater detail being given to development of the CSM. Should this latter scenario prove to be the case, rather than taking many more samples, effort could potentially be better spent on reviewing historic data, in order to refine the CSM. This could well identify potential ‘hot spots’ which could be sampled, delineated and removed.

B4 How to decide which example your dataset is closest to

To decide which data type your dataset is closest to, follow these rules of thumb in the order listed.

1. Your data is FAT-TAILED if all of these conditions apply:
   a. All samples are greater than zero (using an appropriate non-detect rule) and can be transformed to a log scale.
   b. The histogram of the actual values consists of a main body containing most of the values plus a few extreme values.
   c. The histogram of the log values is clearly and obviously not symmetric.
   d. The mean value is greater than the upper quartile of the actual values (see paragraph below).

2. Otherwise your data is LOG-SYMMETRIC if all of these conditions apply:
   a. All samples are greater than zero (using an appropriate non-detect rule) and can be transformed to a log scale.
   b. The histogram of the actual values is clearly and obviously not symmetric.
   c. The log values of the data is considered symmetric as per the conditions for the symmetric data in step 3 below.
   d. The mean value is less than the upper quartile of the actual values (see paragraph below) but clearly greater than the median of the actual values.

3. Otherwise your data is SYMMETRIC if all of these conditions apply:
   a. The mean value is close to the median value.
   b. The shape of the histogram below the median value is an approximate mirror image of what can be seen above the median value i.e. it is roughly symmetric.

4. Otherwise consider your data to be SYMMETRIC even if it does not conform with the symmetric definition above.
   a. For example, a dataset where the mean value is clearly less than the median value is unlikely to be truly symmetric as per definition 3. An example of when this occurs is if your data is negatively skewed.
   b. However, the sample size requirements for such data will be the same as for a symmetric dataset hence why for the purposes of this guidance, any dataset that is not fat-tailed or log-symmetric can be considered symmetric.

All of these points bar one has already been flagged in sections B1 to B3. The one new point appears in points 1d and 2d. Point 2d states that if the mean value is less than the upper quartile whilst also being greater than the median, then you probably have log-symmetric data. Conversely, point 1d states that if the mean value exceeds the upper quartile, you probably have fat-tailed data. You should be able to see that the three statistics, mean, median and upper quartile form a sliding scale on which the likely data distribution can be identified. This sliding scale becomes more apparent when plotted as a box plot as explained in section 4.2.

These values (in mg/kg) are listed in ascending order below for each example dataset described in this appendix.

- Symmetric: Mean (1.16) < Median (1.17) << Upper Quartile (1.39)
- Log-Symmetric: Median (0.5) < Mean (0.75) < Upper Quartile (0.91)
- Fat-Tailed: Median (0.03) << Upper Quartile (0.35) < Mean (0.5)

Appendix C – Glossary of basic statistical and mathematical terms

The following functions in Microsoft Excel will allow you to reproduce the calculations given in this document. All references to CELL or RANGE refer to the cell range where your sample data is held e.g. cells A1:A50.

- **LN(cell)** calculates the natural logarithm of a cell.
- **COUNT(range)** counts how many numerical values there are in a range.
- **AVERAGE(range)** calculates the mean value.
- **STDEV(range)** calculates the standard deviation. Do not use the alternative STDEVP() function!
- **TINV(confidence level, sample size – 1)** calculates the T statistic for confidence interval calculations.
- **MEDIAN(range)** calculates the median value.
- **MAX(range)** and **MIN(range)** calculate the maximum and minimum values.
- **PERCENTILE(range, K)** calculates the K'th percentile where K is a number between 0 (for 0%) and 1 (for 100%). Note that the minimum is the 0th percentile, the lower quartile is the 25th percentile, the median is the 50th percentile, the upper quartile is the 75th percentile and the maximum is the 100th percentile so PERCENTILE can be used instead of MAX, MIN, QUARTILE and MEDIAN.

References and further information

- American Society for Testing and Materials (ASTM), 2014. Standard Guide for Developing Conceptual Site Models for Contaminated Sites. ASTM E1689-95.
- American Statistical Association (ASA), 2016. ASA Releases Statement on Statistical Significance and P-Values, 7th March 2016. Available at: https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf.
- ASA Statement on P-Values & Statistical Significance: Development & Impact, Keynote speech given on 8th September 2016 at Royal Statistical Society's conference. Available at: https://www.youtube.com/watch?v=B7mvbOK1ipA.
- British Standards Institution (BSI), 2018. Soil Quality – Sampling, BS ISO 18400-104:2018.
- British Standards Institution (BSI), 2020. Soil Quality – Conceptual Site Models for Potentially Contaminated Sites. BS EN ISO 21365:2020.
- Environment Agency, 2000. Secondary Model Procedure for the Development of Appropriate Soil Sampling Strategies for Land Contamination Development. R & D Technical Report P5-066/TR. Available at: https://webarchive.nationalarchives.gov.uk/20140328173022/http://cdn.environment-agency.gov.uk/sp5-066-tr-e-e.pdf.
- Environment Agency, Natural Resources Wales, Scottish Environment Protection Agency and Northern Ireland Environment Agency, 2018. Guidance on the Classification and Assessment of Waste Technical Guidance WM3. 1st Edition v 1.1. Available at: https://www.gov.uk/government/publications/waste-classification-technical-guidance.
- Hughes, W. and Lavery, J., 2014. Critical Thinking: An Introduction to the Basic Skills, 7th ed. Broadview.
- Wasserstein, R.L. and Lazar, N.A., 2016. The ASA statement on P-values: Context, Process and Purpose, The American Statistician, Vol 70, No 2, 129-133. Available at: https://doi.org/10.1080/00031305.2016.1154108.
- Wasserstein, R.L., Schirm, A.L. and Lazar, N.A., 2019. Moving to a world beyond P<0.05, The American Statistician, Vol 73, Issue Sup 1. Available at: https://doi.org/10.1080/00031305.2019.1583913. Note this reference contains links to a large number of articles under the same heading published by other authors. The reader is recommended to start with the above reference before moving on.